KNOWLEDGE DISCOVERY IN BIOMEDICAL RESEARCH AND DRUG DESIGN: THE DEVELOPMENT AND APPLICATION OF BIOLOGICAL DATABASES

JI ZHI LIANG
(M.Sc. NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF SCIENCE

DEPARTMENT OF COMPUTATIONAL SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2003

ACKNOWLEDGEMENTS

With a deep sense of gratitude, I wish to express my sincere thanks to my supervisor, Professor Chen YuZong, for his immense help in planning and executing my research in time. His profound knowledge and kind guidance let me know the process of research, and his valuable suggestions ensure my works carrying on in the right way.

I wish I would never forget our BIDD group. In particular, I specially thank: Dr Cao ZhiWei, Dr Chen Xin, Mr Han LianYi, Ms Sun LiZhi, Mr Wang JiFeng, Ms Yao LiXia, Mr Yap ChunWei and our research staffs: Dr Cai CongZhong, Dr Li ZeRong, and Dr Xue Ying. Without their helps, this work can not be properly finished.

I also wish to thank all friends and colleagues in/out of Dept. of Computational Science. It is them who make my studying and researching life smoothly and joyfully.

Needless to say, I will thank my wife. Without her accompany and encourage, I don't know how far I can go.

I will miss the people, the time and the place forever.

TABLE OF CONTENTS

SUMMARY

The biomedical data grows dramatically year-by-year. Especially with the completion of sequencing by the Human Genome Project, the biological research enters the postgenomic era. To well manage and use these fast-growing data, a large number of biological databases are created as well as various data analysis tools. In this work, studies have been focused on the development of biological databases and their applications in biomedical research and drug discovery.

The development of database is a complex and time-consuming process. The entire process is carried out stage by stage, from data preparation, database construction, to database representation. Different technologies are used in different stages of database development, e.g. information retrieval (IR) and text mining (TM). Following the strategy of database development, two biological databases were developed in this work: the Drug Adverse Reaction Target database (DART) and the Kinetic Data of Biomolecular Interaction database (KDBI). DART collects the literature recorded protein targets that are able to induce, directly or indirectly, the adverse drug reactions (ADRs). Efforts have been made to gather the related information such as the physiological function of each target, binding drugs/agonists/antagonists/activators/inhibitors, corresponding adverse effects, and type of ADR induced by drug binding to a target. This work has been published in the international journal *Drug Safety* [Ji *et al.*, July 2003]. KDBI was created which aims at providing experimentally determined kinetic data of bio-molecular interaction such as protein-protein and protein-nucleic acids described in the literature. Such information is important for mechanistic investigation, quantitative study and simulation of cellular

processes and events. This work has been published in the international journal of *Nucleic Acids Research* in 2003 [Ji *et al.*, January 2003].

In addition to simply providing the information, further analysis on these two databases was made. Two knowledge discovery applications of the DART database were investigated. One of them intended to identify the ADR targets based on protein primary sequences using the learning algorithm of Support Vector Machines (SVMs). A model was constructed, trained and optimized using known ADR targets of DART database as positive data. The optimized model was later able to classify the potential ADR targets and non-ADR targets. Similar work of protein family classification using SVM was published in *Nucleic Acids Research* [Cai *et al.*, 2003]. The knowledge discovery of DART database was also made to facilitate drug discovery. In this work, the potential ADR targets of 11 marketed AIDS drugs were predicted by searching the DART database. The prediction involved a docking software INVODOCK, which is able to optimize the drugs docking into the proteins by searching the protein cavity database. For each studied drugs, the docked proteins were listed. They are the possible targets while the drug is admitted to the body. These proteins include the potential therapeutic targets, ADME (Absorption, Distribution, Metabolism, Excretion)-associated proteins, and ADR targets. A good way to identify these targets is searching the respective target databases. For example, by searching the drug adverse reaction targets database DART, one can easily figure out whether the studying drug is safe enough and what kinds of adverse effects it may induce. Respective target databases for therapeutic targets [Chen *et al.*, 2002] and ADME-association proteins [Sun *et al.*, 2002] were constructed previously with the effort

of our group members. Finally, a databases-supported Computer-Aided Drug Discovery system (CADD) was established and studied.

The knowledge discovery of kinetic database KDBI was also studied by the construction of protein-protein interaction network. Comparing to other similar networks available on-line, all of the protein-protein interactions in the KDBI are confirmed by the literature with kinetic value. Such protein-protein interaction network facilitates biological pathways study both in quantity and quality. It is also helpful for the identification of new therapeutic targets, even drug discovery. The network is still preliminary and will be extended and consolidated with more new data added in.

CHAPTER 1 INTRODUCTION

1.1 History of Database Technology

Database and Database Management System (DBMS) is one of the most important classes of modern information technology. The term “data base” is thought to be adopted first by the SDC, the Rand Corporation group around 1960, which described the shared collection of information on which all these views were based [Haigh *et al.*, 2003]. The development of the first database was involved as part of the famous SAGE anti-aircraft command and control project, which was the first major system able to respond immediately and directly to representations of various information to all users. This requires the management of central, electronic and instantly accessible file of enormous size. As a result, such system was invariably written in low-level assembly language in mid-1960s, when few practical tools were available for use in the construction of a database. However, by that time, the concept of management system of database was not formed yet. Until 1968, the term “data base management system” was standardized by the Data Base Task Group (DBTG), by combining two previously separated concepts: the formerly vague “data base” itself and the well defined “file management” or “information storage” software. The acceptance of the DBMS concept implicitly redefined the “data base”, which became a new, narrower and much clearer idea. At present, data base is an integrated collection of data, usually stored on the secondary storage devices such as disks or tapes, and maintained by DBMS.

The application of databases is broad, both in the academia and industries. This thesis reports our research on the development of biological databases and their applications in drug discovery and knowledge discovery in specific areas of biomedical science. The relevant technologies of database development and knowledge discovery are discussed as well.

1.2 Development and Categories of Biological Databases

1.2.1 History of biological databases development

In the early days, when a database containing 200 entries of nucleic acids sequence was opened for public access [Dayhoff *et al.*, 1980], the general opinion was doubtful regarding the ability of biological databases to aid in biomedical research. Now, it becomes a routine procedure for the researchers to search specific biological databases to address some questions before expensive experiments are carried out. The latest database issue of *Nucleic Acids Research* lists about 400 different databases covering diverse areas of biological research [Baxevanis *et al.*, 2003] including primary sequence, genetics, intermolecular interactions, pathways, pathology, proteomics, structure and medical information.

The increase is not only in the number of the databases, but also in their size and complexity. Today, biological databases can be huge in size as the large-scale primary archiving projects, such as GenBank and SWISS_PROT. For example, the major protein database SWISS_PROT contains 12,7863 entries as of June of 2003. In each entry, a variety of information is included, for example, protein name, synonym, gene name,

organism species, primary sequence, taxonomy cross-reference, physiological function, domain. and many cross-links to other databases. Furthermore, to easily access a database, a powerful searching engine is provided for keyword or ID search, as well as some useful Bioinformatics tools such as sequence alignment. Facing the ever-increasing data, flat files database management systems, which were used for storage and representation of data by databases of early ages, are no longer sufficient for the present day biological databases. The more powerful and functional database management systems such as the *Relational Database Management Systems* (RDBMS) are in demand to efficiently maintain the comprehensive and cross-related information stored in databases. At the same time, internet technologies such as World Wide Web (WWW) and visualization technologies are acquired to make the representation of databases more user-friendly. Recently, there appears to be a trend for the traditional databases to evolve into knowledge bases. Therefore, various knowledge discovery technologies have been developed and employed, that will be discussed in a later section.

1.2.2 Categories of biological databases

Today there are a large number of databases available on-line ranging from the large–scale project archives such as SWISS-PROT to individual, specialized collection such as Receptor Database [Nakata *et al.*, 1999]. According to the scope of databases, a biological database can be grouped into three categories [Frishman *et al.*, 1998]:

General biological databases, which store the raw data of DNA/protein sequence, structure, and biological and medical literature. Examples include: the nucleic acid and

protein primary sequence databases such as GenBank [Benson *et al.*, 1999] by National Center of Biotechnology Institute (NCBI), Nucleotide database of European Molecular Biology Laboratory (EMBL) [Stoesser *et al.*, 1998], and DNA Data Bank of Japan (DDBJ) by the National Institute of Genetics (NIG), Japan [Tateno *et al.*, 2002]; the protein databases such as Protein Knowledgebase SWISS-PROT/TrEMBL [Bairoch *et al.*, 2000] by Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI), Protein Information Resource (PIR) by Georgetown University Medical Center (GUMC), USA [Wu *et al.*, 2002]; the original structure databases such as Protein Data Bank (PDB) by Rutgers, The State University of New Jersey, USA [Sussman *et al.*, 1998], the Structural Classification of Proteins database (SCOP) by Medical Research Council (MRC), Cambridge, UK [Murzin *et al.*, 1995]; the biological and medical literature databases like MEDLINE by NCBI [Wheeler *et al.*, 2003]. Databases of this category are repositories of original experimental results. They are normally huge in size and operated by some well-known large research institutes, however, there are also some comparatively small databases in this category such as the searchable database of multidimensional biological images, BioImage by EBI [Carazo *et al.*, 1999]. Sometimes international collaborations of research institutes help to standardize and enrich the databases. The typical such cooperation is the International Nucleotide Sequence Database Collaboration among GenBank, EMBL and DDBJ (Figure 1.1). Generally, databases of this category are a basis for other databases, bioinformatics tools and commercial software.

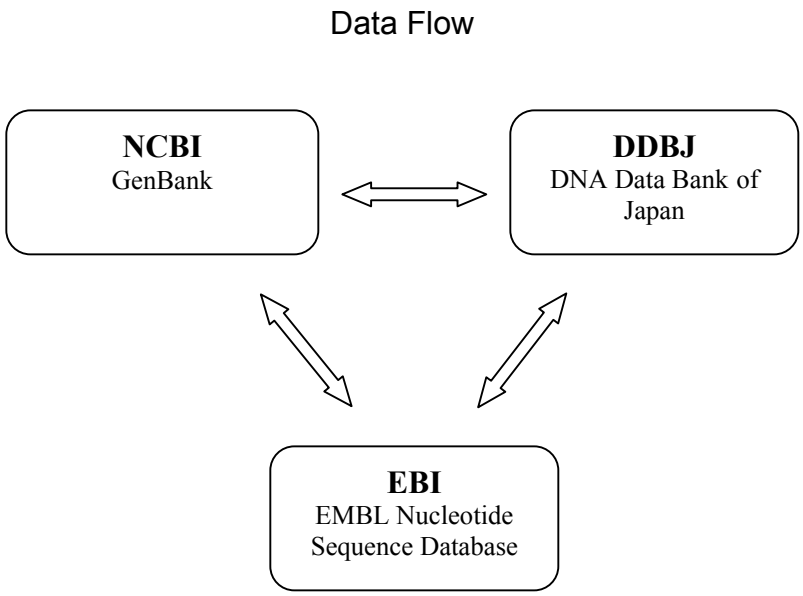
Derived databases, whose data are derived from the general biological databases, but that, contain novel information. For example, the database of protein families and domains (PROSITE) consists of biologically significant sites, patterns and profiles that help to

reliably identify to which known protein family (if any) a new sequence belongs [Bairoch *et al.*, 1994]; the protein families database (Pfam) is a large collection of protein multiple sequence alignments and profile hidden Markov models based on protein primary sequence databases [Bateman *et al.*, 2002]. Databases of this category generate their novel information by analyzing or mining the primary sequence, structure of nucleotides or proteins. The generation process is normally through certain Bioinformatics software or algorithms such as multiple sequence alignments automatically working on the large volume of raw data. Databases of this category regenerate novel information regularly when the respective raw data source is updated.

Subject-specialized databases, which collect individual, specialized information for communities with particular interests. Databases of this category can include databases with original experimental data or derived databases that are based on the general databases. The characteristics of these databases are: subject-specialized, compact in size, and comprehensive in converting their respective subject. The examples include the protein specialized databases: the Comprehensive Enzyme Information System (BRENDA), developed at the Institute of Biochemistry at the University of Cologne that mainly collects enzyme functional data [Schomburg *et al.*, 2002]; Another enzyme nomenclature database (ENZYME) also provides similar information, which is maintained by SIB [Bairoch *et al.*, 2000]; the G-protein coupled receptor database (GCPRD) collects, combines, validates and disseminates heterogeneous data on G protein-coupled receptors (GPCRs) [Horn *et al.*, 1998]; the pathways databases: Kyoto Encyclopedia of Genes and Genomes PATHWAY (KEGG PATHWAY) is the primary database resource for the computerized knowledge on molecular interaction networks such as pathways and

complexes [Kanehisa *et al.*, 2002]; the PathDB developed by National Center for Genome Resources (NCGR), USA, is both a data repository and a system for building, visualizing, and comparing cellular networks (http://www.ncgr.org/pathdb/); the gene databases: Transcription Regulatory Regions Database (TRRD) is an informational resource containing an integrated description of the gene transcription regulation [Kolchanov *et al.*, 2002]; BodyMap focuses on human and mouse gene expression that is based on site-directed 3'-expressed sequence tags generated at Osaka University [Sese *et al.*, 2001]; the intermolecular interaction databases: the Biomolecular Interaction Network Database (BIND) archives biomolecular interaction, complex and pathway information [Bader *et al.*, 2003]; the Database of Interacting Proteins (DIP) documents experimentally determined protein-protein interactions [Xenarios *et al.*, 2000]. There are many other subject-specialized databases available for the interests of different communities; for example, our therapeutic target database (TTD) is especially designed for the identification of the therapeutic target proteins documented in the literature [Chen *et al.*, 2002]. Subject-specialized databases make up the major portion of the biological databases, especially, the small and medium size databases. These are functional databases and often able to aid in biological/medical research, drug discovery, and human healthcare.

Figure 1.1. The collaboration of international institutes on nucleotide sequence databases



1.3 Role of Database in Analyzing Biomedical Data

1.3.1 Analysis of biomedical data with databases

At the end of 20th century, with the efforts of some individual genomics companies and the international Human Genome Project groups, the entire human genome has been sequenced. When the applause for this grand achievement is fading, more challenging tasks emerge. The challenges are how to identify the genes and other functional fragment from the vast raw genetic sequence? How to figure out the physiological functions of the proteins or peptides coded by those genes? In the long-term, how to elucidate the "*underlying molecular mechanisms of disease and thereby facilitating the design in many cases of rational diagnostics and therapeutics targeted at those mechanisms*" [Waterston *et al.*, 2002]. To answer these questions experiments alone are not enough, and sometimes beyond reach in the near future. A better solution is to combine experimental data and technologies of informatics to seek the clues, which has introduced a new discipline: Bioinformatics. Biological database technology is one of important area of Bioinformatics. Database organizes biological data in a rational way, which offers a platform for further analysis and knowledge discovery from these data. Development and application of biological databases have pushed and accelerated the development of Bioinformatics as a discipline.

Bioinformatics is *the computer-assisted data management discipline that helps us gather, analyze, and represent biological information in order to understand life's processes* [Persidis *et al.*, 1999]. As described in the Oxford English Dictionary, the definition of Bioinformatics is "*conceptualizing biology in terms of molecules and applying*

'informatics techniques' to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications".

The start of Bioinformatics can be traced back to mid 1970s, when automated protein and DNA sequencing became available. The early application of bioinformatics was typically associated with database of gene/protein sequences, when the databases were accessed locally and with limited analysis tools. With the development of internet technology, in the late 1980s, those databases were also accessible remotely, and more analysis tools became available. From the 1990s on, the popular use of internet and the explosion of biological data, in some sense, has made Bioinformatics equally attractive to academic and company scientists. And because of the efforts of these scientists and funding agencies such as NIH in USA and EMBL in Europe, Bioinformatics became more and more prominent and diverse.

Biomedical data analysis of different levels

In definition, the ability of Bioinformatics is to gather, store, classify, analyze, distribute, simulate, and predict biological information derived from sequencing, functional analysis projects such as protein 3D structure analysis, metabolic pathways simulation, human genes extraction and literature of biological and medical research. The technologies used in Bioinformatics which include databases, different kinds of analysis tools based on sequence, structure and function, drug design assistant system, or data mining (knowledge

discovery) based on databases. According to the aims of these technologies, biomedical data analysis can be roughly categorized into three levels.

At the first level, the biological data is collected and well organized so that users are allowed to access and retrieve the information for further analysis. The most important and typical technology at this level is a database. Data from different source is collected and deposited in respective databases. To well organize the vast, high-dimensional, cross-related data, a good data structure and database management system (DBMS) are desired. The data warehouse technology, and some commercial *Relational Database Management System* (RMDBS) such as *ORACLE* and *SYBASE* are thus adopted. For most of public and commercial biological databases, a user-friendly interface to the databases and internet remote access is also provided, through which the data is distributed worldwide for further data analysis.

Databases are widely used in academic research, therapy support, and therapeutic industry. A good database can reduce aid in research, clinical diagnosis, and new drug discovery. A good example is therapeutic decision-making in stages III and IV head and neck cancer treatment [Gleich *et al.*, 2003]. The cases of head and neck cancer in the patient databases were reviewed and analyzed using the Kaplan-Meier method. It was found that the age, co-morbidity, and advanced stage on survival of patients were closely linked. Thus, the site and stage-specific treatment based on the data in the databases would be useful in counseling patients with advanced head and neck cancer. Searching databases for answering specific questions has become a routine practice for most researchers. This trend has brought up the tide of development of databases and the analysis software based

on the databases in recent years. Other than the well constructed databases, much information on-line is simply listed in flat files or tables. These web pages or tables are commonly specialized on certain topics. They are more focused though they may be small in size and limited in the completeness of information. One example is the page of *PROLYSIS* on the protease and protease inhibitors at (http://delphi.phys.univ-tours.fr/Prolysis/index.html). Another typical example is the page of Tools for Glutamate Receptor Research by University of Bristol at (http://www.bris.ac.uk/synaptic/info/tools.html), which details agonists and antagonists for NMDA, AMPA/Kainate and mGlu receptors.

Once the data is made available, an analysis of these data becomes possible. At the second level of Bioinformatics, a number of data analysis tools are developed. These tools use the raw data or derived data of DNA/protein sequence, structure, and literature information to generate new information. For example, sequence alignment tools *FASTA* [Pearson *et al.*, 1988] is able to search DNA/Protein sequence databases, evaluate similarity scores, and identify periodic structures based on local sequence similarity. Similar tasks can also be done by *BLAST* [Altschul *et al.*, 1997]. Other tools include translating nucleic acid sequence to peptide; protein identification and characterization; pattern and profile searches; primary structure analysis. A list of such tools can be found in *ExPASy* Proteomics tools page (http://tw.expasy.org/tools/), which are free for researchers. EMBL-EBI Toolbox also collects different categories of tools for the fields of Bioinformatics (http://www.embl-ebi.ac.uk/Tools/index.html). Comparing to the free tools on-line, some Bioinformatics companies develop commercial Bioinformatics software of more functions and abilities. For example, the molecular modeling software *SYBYL* developed by

TRIPOS is a program able to build, study and manipulate molecules including macromolecules like nucleic acids and proteins. It also provides some powerful tools for molecular dynamics, energy minimization, homologous modeling. Special hardware, e.g. SGI graphic workstation, is required to ensure the program work properly. Similar commercial software of Bioinformatics is *INSIGHT II* developed by ACCELRYS.

Bioinformatics tools such as sequence alignments, pattern searches are able to analyze the raw data, thus to summarize the useful rules or information, even to simulate protein structure or the biological systems such as metabolic pathway. However, some tools for the analysis, calculation, and simulation may be inadequate for the practical application such as the pharmaceutical industry. Extracting the hidden meaningful information from the data pools and further predicting the new events in advance is expected. For example, how to identify the individual genes from the DNA sequence? How to predict the protein structure based on the sequence? How to predict protein/protein or protein/ligand interactions? Fortunately, the introduction of new knowledge discovery technologies and algorithms make these attempts possible. A good example is the application of data mining technologies such as SVM, decision trees in gene identification [Rosenquist *et al.*, 2001], protein/protein interaction prediction [Bock *et al.*, 2001] and therapy support [Dusseldorp *et al.*, 2001]. These approaches are not yet mature, and more new technologies and algorithms are being introduced to further improve them. More about data mining will be discussed later.

In conclusion, the flood of biological data has catalyzed the construction of databases for the data storage and distribution. It has also stimulated the development of respective data

analysis tools and software. The Bioinformatics tools/software are applied in life science research [Boguski *et al.*, 2003], medical research [Lynn *et al.*, 2003], therapy decision-making [Sarachan *et al.*, 2003], pharmaceutical industry [Liebman *et al.*, 2002] and many other biological relevant fields. For example, support vector machines (SVMs) software was used to analyze the microarray expression data thus classify and validate the cancer tissue samples from normal tissue samples [Furey *et al.*, 2000]. Many new molecular-based technologies such as Genomics, Proteomics, transcriptional profiling, gene expression patterns and respective software have been applied in new drug discovery. The complete genome sequence information of human, bacteria, and virus, with subsequent bioinformatics analytic tools may support computer-aid drug design [Haney *et al.*, 2002]. The databases and Bioinformatics software is developed for different purposes; however, it is widely acknowledged that the long-term value or final object of Bioinformatics is not the development or use of tools, but knowledge discovery so as to improve the human health.

1.3.2 An example: database for kinetic study of biomolecular interactions

Proteins and nucleic acids can be regarded as one of the basis of the modern molecular biology. Almost all the biological events involve proteins or nucleic acids. The study of biological events is the way for us to understand human body behavior, possible etiology and therapy. Such study can be carried out in three progressive stages: first is the physiological function of individual molecule itself, second the interaction between the bio-molecules, and finally the cellular process composed of different bio-molecular

interaction. The discovery of physiological functions of biomolecules is normally by repeating experiments such as catalyzing analysis and binding analysis on the respective molecules. Unfortunately, it is costly to try all the analysis to determine the molecular function. An alternative way is through the use of Bioinformatics analysis tools for facilitating function discovery. One can compare the respective protein primary sequence with the sequences deposited in databases such as SWISS_PROT or GenBank by using sequence alignment tools such as *BLAST* and *FASTA*. It is believed that homology in protein primary sequence always indicates similarity in physiological function. The prediction of protein function can be further verified by rationalized and focused experiments. The interaction between molecules, including protein-protein, protein-nucleic acids and protein-ligand, is normally identified by binding experiments and kinetic analysis. The binding analysis confirms the interaction between the molecules, while the kinetic analysis reveals the time course of the interaction. Cellular processes and underlying molecular events involve complex interactions and cross talks between individual molecules, pathways and networks of pathways [Downward *et al.*, 2001; Lengeler *et al.*, 2000]. Simply, the cellular processes or biological pathways are the networks of molecular interactions, which are often used as the clues of etiology and therapy. The distinctive interactions are connected to each other and may affect others. The effects of upstream molecules on the downstream molecules are unequally, however, quite different due to different possibilities of reaction happening. Therefore, quantitative as well as mechanistic understanding of these interactions is important for exploration and engineering of cell behavior and for the development of novel therapeutics to combat diseases. A number of databases of molecular interactions [Bader, 2001; Xenarios, 2002], pathways [Goto *et al.*, 1997; Igarashi *et al.*, 1997; van Helden *et al.*, 2000] and enzyme

reactions [Goto *et al.*, 1998] have been developed. These databases provide comprehensive information about interacting molecules, molecular complexes, pathways, chemical reactions, and conformation changes. The kinetic data for these interactions, important for mechanistic investigation, quantitative study and simulation of cellular processes and events [Sahm *et al.*, 2000; Fussenegger *et al.*, 2000; Haugh *et al.*, 2000], is not provided in the existing databases. Therefore, in this work, a Kinetic Data of Biomolecular Interaction database (KDBI) is developed to provide kinetic information for protein-protein, protein-ligand, and protein-nucleic acids interactions. Furthermore, knowledge discovery from the KDBI database is tried to construct the protein-protein interaction network, which could be part of biological pathways. It is expected that both the kinetic database KDBI and its derived protein-protein interaction network will help to better understand of disease etiology and better therapy.

1.4 Role of Databases in Facilitating Drug Discovery

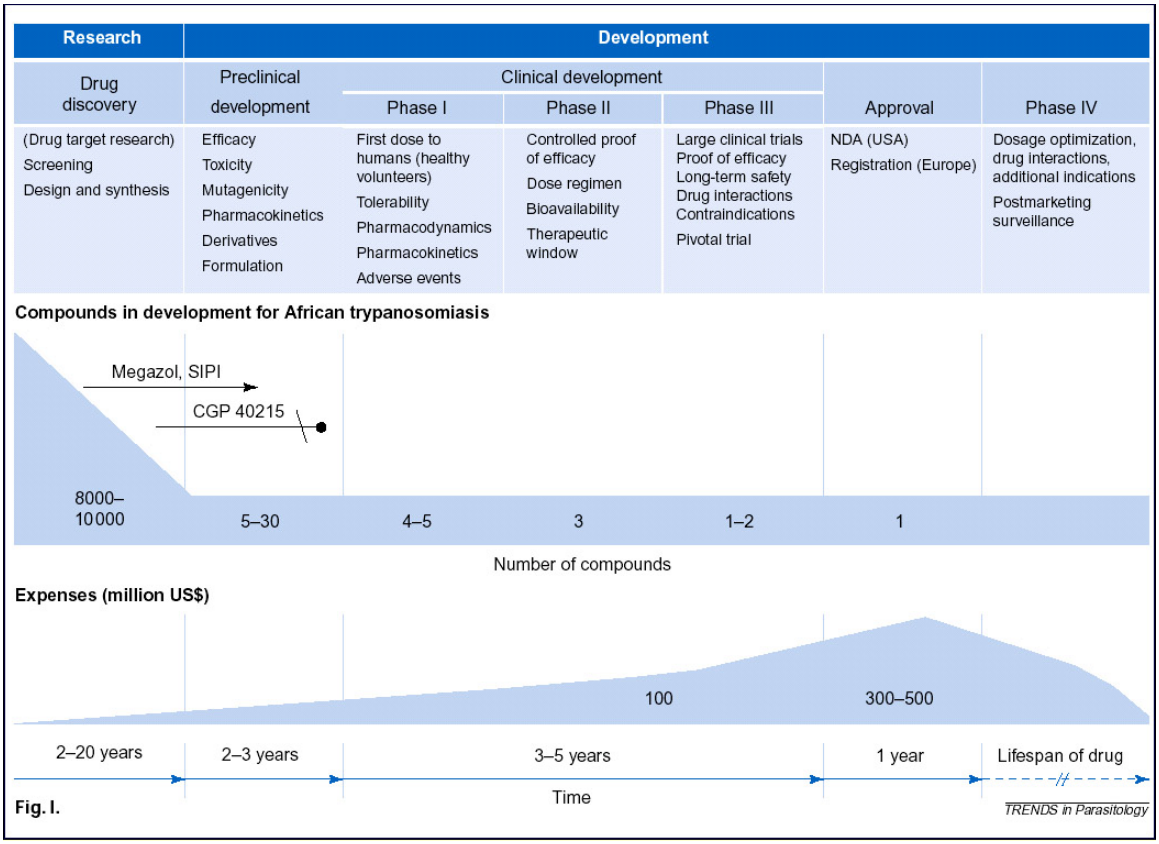
1.4.1 Overview of emerging technologies of drug discovery

Drug discovery is complex and costly process. It is an innovative, creative, and iteratively experimental science, which is more than the application of basic research knowledge and technologies [Black *et al.*, 1986]. It involves many facets of project management and research [Jacques *et al.*, 1992].

Generally, before a drug reaches the market, it needs to go through three main stages: drug discovery and testing in the lab, clinical evaluation, and market feedback (Figure 1.2). Each stage of new drug development is time-consuming and costly, especially the initial

stage of drug discovery, which can last up to 20 years. Thousands of candidate compounds are screened, and only a limited number and success of compounds reach pre-clinical development for their activity, efficacy, selectivity, bioactivity, and pharmacokinetics studies. The pre-clinical development process may take up several years depending on the number of the compounds. Those compounds that fulfill the clinical requirements, normally only few, will be evaluated in further clinical trials. The clinical trials are composed of four phases: Phase I studies determine safety of compound in normal human volunteers using dose-ranging studies. Side effects as well as human pharmacokinetics are established at this stage. Phase II studies involve open-label, single- and multiple-dose studies in the patient population. Efficacy and bioactivity is determined at this stage. Phase III focuses on larger clinical trials proof of efficacy and the establishment of uncommon side effects and drug interactions. Passing these three clinical trials, the drug candidates are eligible to submit to new drug controlling organisms such as FDA for approval of marketing. In the first few years of marketing, the new drugs will still under supervision. The feedback of patients and doctors will be helpful for the dosage optimization, drug interaction and additional indications studies. The normal new drug discovery process is illustrated in Figure 1.2 using new drugs developed for *African trypanosomiasis* as example [Keiser *et al.*, 2001]. The extremely high cost and the long research period makes the development of new drug more and more difficult. Therefore, reducing the costs and shortening the new drug development time would be a stimulator for the pharmaceutical industry.

Figure 1.2. The process of new drug development for *African trypanosomiasis*.



The history of drug development can be traced back to hundreds of years. However, modern drug discovery beginning from early 1940's is mostly based on the synthetic chemistry, biology, biochemistry and pharmacology. The process of drug discovery was often dependent on natural sources and serendipity [Sneader *et al.*, 1990]. A second phase of drug discovery began with the advances in enzymology and biochemistry. At this phase, designing drugs directly interacting with the distinct molecular target became possible. With the increase in computational power, drug discovery has entered a new phase of computer-driven drug discovery, *Computer-Aided Drug Design* (CADD), or *Computer-Assisted Molecular Design* (CAMD). At this stage, further acceleration of drug discovery really becomes possible.

Computer-Aided Drug Design (CADD) is a relatively new technology developed in the last decades. The development of CADD went in parallel with increase in computational power of computers. The demand of high computational power is because of the large calculation of the electronic properties of molecules, which is the foundation of the CADD. According to the focus of studying target/ligand interactions, the CADD approaches can be summarized into the following three groups:

(1) Approaching the problems from the drug perspective with knowledge of the compound structure activity relationship (SAR) within series of pharmacophores. This approach is based on an assumption that the protein and ligand have limited degrees of flexibility [Saunders *et al.*, 1989; Sim *et al.*, 2002].

(2) Approaching the problems from the receptor or enzyme perspective with knowledge of the structure of the receptor or enzyme. The knowledge of the structure and 3D confirmation of the protein target provides an opportunity to identify the amino acid sequences and conformations that are responsible for ligand recognition and efficacy [Fritz *et al.*, 2001; Wheatley *et al.*, 1998].

(3) Approaching the problems with information regarding the receptor/ligand, enzyme/substrate interaction derived by 2- or 3-D structural protein analysis methods. Compared to methods of previous categories, methods of this category pay more attention on the interaction between two molecules. The structural analysis methods include the nuclear magnetic resonance (NMR), X-ray crystallographic, and other methods. The impact of nuclear magnetic resonance (NMR) spectroscopy on rational drug design has recently increased through the description of the so-called structure-activity relationships (SAR) by NMR technique. The analysis of protein structures determined with minimal structural information by NMR can be extended with a particular interest in the utility of these structures for a structure-based drug design program [Huang *et al.*, 2000; Wender *et al.*, 1999].

Looking at the detailed approaches and technologies, the most popular CADD technologies include structure-based approaches and quantitative structure-activity relationship (QSAR) approaches. The structure-based approaches attempt to design drugs in respective of based on the known protein structures, for example, the design of the HIV RT inhibitors based on the known HIV reverse transcriptase structure [Tantillo *et al.*,

1994]. Structural methods such as X-ray crystallography or NMR technology have also been used to study inhibitor-target interactions for antitumor drugs design [Denny *et al.*, 1994]. When no experimental structure is available for the protein target, structures modeled by homology model are used to facilitate drug design [Teeter *et al.*, 1994]. The docking approaches are a series of special structure-based approaches, which use computers to simulate the docking process of ligands to their protein receptor. Various docking approaches have been developed in recent years along with the increase of computational power. Different algorithms have been applied to more properly model the docking process and facilitate drug design [Krumrine *et al.*, 2003]. In cases where the structure of the target protein is unknown and a modeled structure is difficult to derive, it is impossible to use the structure-based drug design. Rather, statistical learning based methods such as the QSAR approaches are applied. QSAR methods attempt to correlate biological activity with physical-chemical properties and structures of molecules. An example of its application is the successful design of the inhibitors for the HIV-1 protease by QSAR [Oprea *et al.*, 1994]. In recent years, applications of QSAR in drug discovery have become supported by QSAR databases [Hansch, 1995; http://mmlin1.pha.unc.edu/~jin/QSAR/].

1.4.2 The need of drug target databases for drug discovery

Drug discovery and development is a complicated and long-term process. It is noticed that knowledge of protein targets of drugs (those proteins to which drugs bind and produce specific effects) play a crucial role in the disease etiology studies, pharmacokinetics studies, toxicity studies. Identification of these target proteins facilitates the design of

drugs with enhanced efficacy and reduced side effects that offer better treatment options for patients. In this work, the adverse drug reaction target database (DART) is created to facilitate the identification of potential toxicity targets to filter out the serious toxicity inducing drug candidates. It is expected that a series of target databases like DART are useful in facilitating rational drug discovery. Three kinds of target proteins are important for drug discovery: therapeutic targets, ADME (absorption, distribution, metabolism, and excretion) associated proteins, and adverse drug reaction (ADR)/toxicity targets.

The proteins to which drugs specifically bind and elicit therapeutic effects are called "therapeutic targets". Diseases are often caused by irregular inhibition or activation of certain proteins in biological pathways. The function of the drugs is to bind to specific proteins in the pathways to re-balance these pathways. Theoretically, all proteins in the pathways could be the potential targets of drugs. However, practically, only those that play essential roles in the pathological pathway regulation will be considered. Even under these circumstances, the selection of targets still needs to be prudent. For example, drugs should only act on pathological pathways but not on pathways controlling normal physiological functions; the selected target proteins should be sufficiently sensitive so that only small amount of drugs are needed to cause curative effects, which thus avoids the possible side effects due to the high dosage of drugs. A practical solution is to collect and study all the existing clinical and experimental therapeutic targets for different diseases. It is estimated that there are approximately 500 therapeutic targets [Drews *et al.*, 2000], the majority of which have been collected by our Therapeutic Target Database (TTD) [Chen *et al.*, 2003].

The metabolite process of the drug candidates, from their intake until their excretion from the body, is important for the efficacy and bioactivity study. This process includes the absorption, distribution, metabolism, and excretion (ADME) of the drugs. Absorption is the process of the intake of drugs into the vascular system. Some drugs are small enough to directly absorb from the gastrointestinal system or other tissues into the blood stream; however, some need the assistance of the transporting proteins. The drugs in the blood stream will be delivered to the pathological tissue with the help of transporters/carriers. Some special transporters/carriers will even bring the drugs to the target proteins so that the drugs can then bind to the therapeutic target and cure the diseases. However, some drugs do not directly interact with their targets. These drugs will be metabolized and their products are the real agents to take effect. The metabolism process involves some particular protein families such as cytochrome P450s. The metabolites of drugs and the remaining drugs will be excreted out of the body with the help of some proteins. The deposit of drugs or their metabolites is one of the causes of the cytotoxicity. Therefore, a successful drug candidate should be absorbed and delivered to their target proteins, to be efficacious, whereas excess compounds should be easily removed from body so as to reduce the side effects. ADME-Associated Proteins database (MADE-AP) gathering such ADME associated proteins surely will be very helpful to identify those drug candidate with practical high efficacy and low toxicity [Sun *et al.*, 2001].

A successful drug should possess both high drug efficacy and low toxicity. The toxicity of the drug, or so-called drug adverse effect, is a major cause for the failure of drugs. The mechanisms leading to the induction of adverse drug reaction (ADR) are diverse. The drugs bind not only the therapeutic targets but also other proteins in the non-pathological

biological pathways; the drugs may irreversibly bind to the therapeutic targets due to the high dosage or their binding ability; the drugs or their metabolites may be deposited in the tissues, and the deposition disturbs the environment of cell such as pH environment and ion gradients and thus lead to the toxicity. Many factors are involved in the ADRs and often related to certain proteins. To systematically study the mechanisms of the ADRs and reduce the possible ADRs during drugs discovery, it is necessary and meaningful to collect all the proteins inducing, directly or indirectly, the ADRs. Therefore, in this work, a drug adverse reaction targets database (DART) is created to collect such ADR target proteins.

1.4.3 Adverse drug reaction target database for drug safety evaluation

All drugs can produce harmful as well as therapeutic effects. As the definition of the World Health Organization (WHO), adverse drug reaction (ADR) is "*any noxious, unintended, and undesired effect of a drug, which occurs at doses used in humans for prophylaxis, diagnosis, or therapy*." This excludes therapeutic failures, intentional or accidental poisoning or drug abuse, and adverse effects due to errors in administration or compliance. The forms of ADRs vary from a single physiological/biochemical parameter to multiple organ failure. According to the clinical perspective, the ADRs can be classified as following [Park *et al.*, 1994]:

Type A: These reactions are predictable in terms of the known pharmacology of the drug and are usually dose dependent.

Type B: These reactions are unpredictable from knowledge of the basic pharmacology of the drug and do not show any simple dose-response relationship.

Type C: These reactions are associated with long-term drug therapy.

Type D: These reactions are due to the delayed effects.

The majority of the ADRs in human are of pharmacological nature. It is estimated that about 75% of the ADRs are type A adverse reactions, which are dose-dependent and normally reversible. It is believed that all the drugs may cause dose-dependent adverse effects. This type of ADRs is predictable, and can sometimes be reduced or even removed when the drugs dosage is reduced or drug treatment discontinused. In contrast to type A adverse reactions, type B ADRs lack correlation between the dose and the toxicity. They are often serious and sometimes even lead to death. Fortunately, this type of ADRs is rare.

The adverse effects induced by drugs are dangerous. They hinder the cure of patient, and they are also the causes of many instances of morbidity. Therefore, the understanding of the possible mechanisms of ADRs would be helpful for the successful treatment of patients. The cause of adverse drug reactions often result from interaction of a drug or its metabolite with either its main therapeutic target or other protein and nucleic acid targets important in the normal cellular functions [Pumfor *et al.*, 1997; Wallace *et al.*, 2000; Park *et al.*, 2000; Rang *et al.,* 1999; Klaassen *et al.*, 2001; Baynes *et al.*, 1999]. Identification and characterization of these adverse effect related protein or other molecular targets constitutes a major focus of pharmacology and toxicology research [Klaassen *et al.*, 2001; Kong *et al.*, 1999; Monks *et al.*, 1998]. Knowledge about these targets not only facilitates the study of the mechanism of ADR, it has also been widely used in the development of

experimental techniques and computer tools for molecular analysis and high-throughput screening of ADRs as an early risk assessment tool [Gerhold *et al.*, 1999; Nuwaysir *et al.*, 1999; Barratt *et al.*, 1998; Chen *et al.*, 2001]. Rapid advance in genetic [Peltonen *et al.*, 2001], structural [Sali *et al.*, 1998] and functional [Koonin *et al.*, 1998] genomics is providing increasingly more comprehensive information about adverse effect related genes, proteins and pathways. This helps to broaden the scope of drug safety evaluation R&D to include such tasks as analysis of pharmacogenetic implication of sequence variation or expression pattern alterations of adverse effect targets [Smith *et al.*, 2001; Pirmohamed *et al.*, 2001; Vesell *et al.*, 2000].

Traditionally, knowledge about known ADR targets is extracted from literature search, which can be time consuming and difficult particularly for non-expert. Therefore, a publicly accessible database with comprehensive information about these targets provides a convenient and useful platform for obtaining relevant information. The information of particular interest includes the functional aspects of ADR targets, mode of interaction of a target with binding drugs and ligands, as well as the adverse effect due to the binding of a drug or a chemical to each target. To the best of our knowledge, such a publicly accessible database is not yet available. Thus, we construct a Drug Adverse Reaction Target (DART) database, which contains information about the literature-described known targets related to adverse effects of drugs [Ji *et al.*, 2003].

1.5 Databases Knowledge Discovery

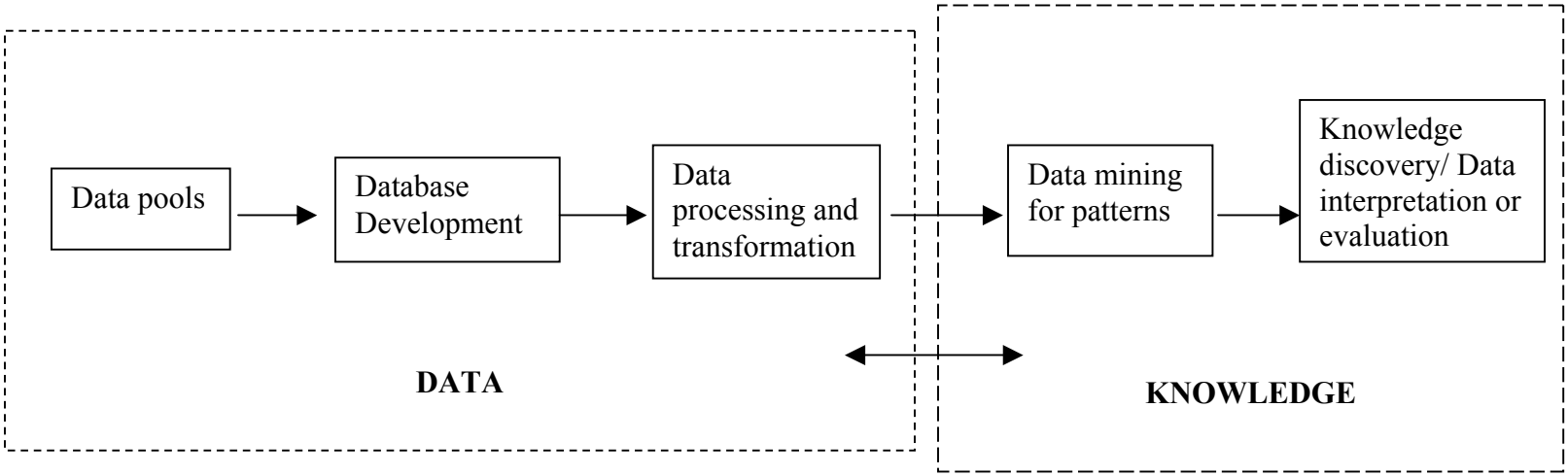
1.5.1 Key role of data mining in evolution of data bases into knowledge bases

Today, databases along with their supporting DBMSs are widely used in academic research, business, and industries. Database use grows quickly with the expansion of the internet. It is noticed that the development of the database is not limited to its application in various domains, but also the database itself. Good integration of data, efficient searching and retrieval engine, and convenient but powerful management has become a characteristic of well-constructed databases. However, that is not enough. The final objective of databases is offering useful information, the knowledge, rather than some unrelated plain data. Therefore, databases should present more than data that are to difficult to understand, but the information they contain; in other words, "data base" should evolved into "knowledge base". The evolution of knowledge bases is a variable process. However, it contains one critical step, which is the knowledge discovery from the databases. In Figure 1.3, the process of database evolution is illustrated. The data deposited in the databases is transformed and mined for patterns using different knowledge discovery technologies. The patterns are further interpreted as knowledge. Thus, the "data base" successfully evolved to "knowledge base". During this evolution process, data mining plays an important role.

Data mining, sometimes also called *Knowledge Discovery in Database* (KDD), has been defined as "*the nontrivial extraction of implicit, previously unknown, and potentially useful information from data*" [Frawley *et al.*, 1992]. It is a powerful new technology with great potential to extract the most important information from data warehouses. Data

mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Its excellent classification and pattern extraction abilities have been recognized and data mining has been adopted to generate of novel information during the construction and application of databases, especially in the evolution of database from the “data base” to “knowledge base”.

Figure 1.3. Evolution of database to knowledge base



1.5.2 Data mining technologies for knowledge discovery from biological databases

In recent years, there is an explosive growth of biological data. Understanding knowledge buried in the enormous amount of data deposited in different databases has become a key task in biomedical communities in general and the Bioinformatics community in particular. As a result, increasing efforts have been directed at the application of data mining in the knowledge discovery in various areas of biomedical research. The introduction of data mining in biomedical research in turn enables the development and application of the data mining technology in biopharmaceutical industry (in general, biopharmaceutical application can also be considered as a part of Bioinformatics). It is no surprise that the biopharmaceutical industry uses data mining to process and analyze the enormous amount of diverse biological information. The collected information ranges from annotated databases of disease profiles and molecular pathways to sequences, structure–activity relationships, chemical structures of combinatorial libraries of compounds, and individual and population clinical trial results. The use of traditional statistical data analysis methods faces a difficulty in solving complex relationships between various types of information. The problem becomes increasingly severe since more and more experimental data is becoming available. A similar situation is observed in biological research and the biotechnology industry: the Human Genome Project has sequenced billions of nucleotides; more people are moving into life science and more experimental data of bio-processes is becoming available; new application of biotechnology such as DNA microarray analysis is generating thousands of new data sets; chemists synthesize more and more compound

libraries for drug discovery. Thus, inevitably, new and powerful data analysis method is needed, and data mining is a useful tool for such a purpose.

Because of the complexity and variety of biological events, different approaches of data mining have been developed and used for the specific applications. So far six types of approaches have been developed:

Influence-based mining: complex and granular (as opposed to linear) data in large databases are scanned for influences between specific data sets, and this is done along many dimensions and in multi-table formats. These systems find applications wherever there are significant cause-and-effect relationships between data sets, for example, in large and multivariant gene expression studies, which are basis of areas such as pharmacogenomics [Burge *et al.*, 1997; Iseli *et al.*, 1999].

Affinity-based mining: large and complex data sets are analyzed across multiple dimensions, and the data mining system identifies data points or sets that tend to be grouped together. These systems differentiate themselves by providing hierarchies of associations and showing any underlying logical conditions or rules that account for the specific groupings of data. This approach is particularly useful in biological motif analysis whereby it is important to distinguish accidental or incidental motifs from ones with biological significance [Narasimhan *et al.*, 2002; Jonassen *et al.*, 2002].

Time delay data mining: the data set is not available immediately and in complete form, but is collected over time. The systems designed to handle such data look for patterns that

are confirmed or rejected as the data set increases and becomes more robust. This approach is geared towards long-term clinical trial analysis and multi-component mode of action studies, for example [Bellazzi *et al.*, 1998].

Trend-based mining: the software analyzes large and complex data sets in terms of any changes that occur in specific data sets over time. The data sets can be user-defined or the system can uncover them itself. Essentially, the system reports on anything that is changing over time. This is especially important in cause-and-effect biological experiments. Screening is a good example, where responses over time to particular drugs or other stimuli are being collected for analysis. The software is designed specifically for this purpose and can identify multiple trends very efficiently [Lavrac *et al.*, 1999].

Comparative data mining: it focuses on overlaying large and complex data sets that are similar to each other and compares them. This is particularly useful for all forms of clinical trial meta analyses where data collected at different sites over different time periods, and perhaps under similar but not always identical conditions, need to be compared. Here the emphasis is on finding dissimilarities, not similarities [Nandi *et al.*, 2002].

Predictive data mining: data mining alone is lacking somewhat if it is unable to also offer a framework for making simulations, predictions, and forecasts, based on the data sets it has analyzed. It combines pattern matching, influence relationships, time set correlations, and dissimilarity analysis to offer simulations of future data sets. One advantage here is that these systems are capable of incorporating entire data sets into their

operations, and not just samples, which significantly increase their accuracy. Predictive data mining is used often in clinical trial analysis and in structure–function correlations [Zien *et al.*, 1999].

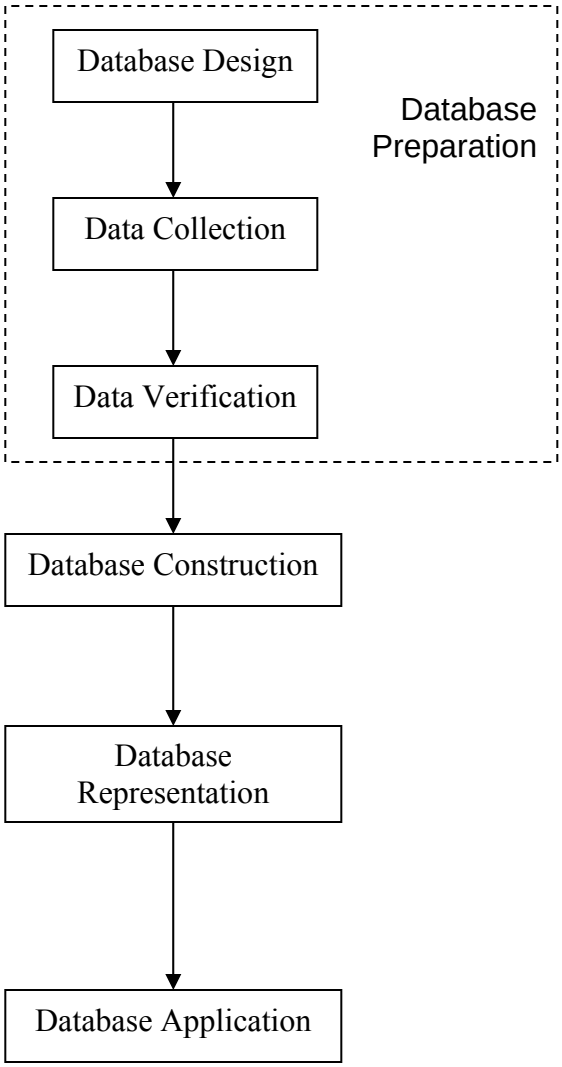
It should be realized that the classification of these six approaches is based on different situations of data analysis, other than specific algorithms of data mining. Actually, all of the data mining methods can be used in biological data analysis, and the approaches may make use of one or more of these methods at the same time. Life science is diverse; therefore, the application of data mining in Bioinformatics also should be diverse. The application of data mining is gradually accepted and applied in different areas. However, there are two critical factors that limit its application: a larger, well-integrated warehouse (databases) and a good understanding of the event that the data mining is to be applied to.

In the coming chapters, we will briefly discuss the strategy of database development (Chapter 2). Following the strategy, two databases: Averse Drug Reaction Target databases DART (Chapter 3) and Kinetic Data of Biomolecular Interaction database KDBI (Chapter 4) were constructed. Applications based on these databases were also carried out to facilitate both biomedical research and drug discovery. In the final chapter (Chapter 5), conclusion is made to previous studies.

CHAPTER 2 STRATEGY OF DATABASE DEVELOPMENT

The development of a database is a complex process of data collection, deposition and representation. During the entire process, different technologies are used, such as data mining and internet communication technologies. Generally, the database is developed in three stages as illustrated in Figure 2.1: database preparation, database construction, and database representation. The database preparation stage is the process of data preparing before they are uploaded into database. It is a complicated and time-consuming, which can be concluded in three steps: Database design, Data collection, and Data verification. Database construction mainly deals with the data deposition related process, especially the selection of database management systems and the practical construction of the database structure. Database representation is able to represent and redistribute the data stored in the database. Normally, internet communication and visualization technologies are intimately involved in this stage. Now we will discuss the development of database stage by stage.

Figure 2.1. Flowchart of database development



2.1 Database Preparation

2.1.1 Consideration of information content and database structure

Database design is the first step and the most critical one in the database development process. It is important because, at this step, one must determine the content and objective of the database, which are the reason and the essence of the existing of database. Therefore, before starting the construction of the database, several questions need to be answered: (1) What information will the database provide? (2) Is the information new, particular and meaningful? (3) Is the information of database comprehensive enough? The first question addresses the need for the objectives of database development. A database is constructed for the purpose of solving or helping to solve problems. As a repository of specific types of information, the usefulness of a database is based on the information stored or interpreted. For example, SWISS_PROT protein database provides the comprehensive information about proteins, and PubMed stores the abstracts of biomedical research literature. To answer the second question, literature survey sometimes is necessary. While searching on-line sources or paper publications, one needs also to seek the answers for: Is there any particular community interested in the database? Is there any database providing similar information? Whether the new database provides more detailed, accurate, comprehensive or extra information compared with the existing databases? After establishing the first two prerequisites, there is one more point that needs serious consideration: what information will be included in the database so that the database can be practically useful? It is generally agreed that the modern databases should be a “knowledge base” rather than “data base”. Providing the novel information alone will

dramatically reduce the usefulness and application of the database; therefore, normally, relevant information is included to make the database comprehensive and applicable.

Designing an efficient data structure is difficult. The difficulty comes from the large number of information concepts and the complex relationship between the concepts. A good method to properly manage diverse information is to use a flow chart. By drawing flow chart, distinct information concepts are allocated into different information blocks. For example, the information about proteins such as protein names, synonyms, gene name will be organized together in one block; and information about ligands such as ligand name, synonyms, chemical formula, etc will put together in another block called ligand information. For those correlated information, whether they are in the same information block or not, lines are used to link them directly. By using the connecting lines, the relationships between distinct information concepts and between different information blocks are properly established. The practical examples can be seen later in the development of databases DART and KDBI. The advantages of using flow chart for database architecture design include: (1) It gives an overview of database components, which makes later data collection relational. (2) It clearly determines the relationship between different information and information groups, which will facilitate later database construction. (3) Future modification or expansion of database will be easier by adding new components and building their connection with other components. (4) Integration with other databases constructed in the similar ways will be feasible and efficient. (5) Teamwork is possible by following the same data structure blueprint.

To conclude, database design is an important stage during the database development. During this stage, the significance of database is determined, and the data structure of database is properly drawn.

2.1.2 Data collection methods

The data collection is the process of gathering information from different biological resources. The resources include the original data of biological findings, experimental results, and clinical researches; the existing biological and medical databases, patents and even newspapers. To acquire information from various resources, different collection methods may apply. The data collecting methods can be as simple as (1) manual collection, and complicated as (2) information retrieval (IR) and (3) text mining (TM). Manual collection of data is the basic method. Normally, the original and novel information from research is collected using this method. Some information scattered in different resources is difficult to extract by programs and also relies on manual work. The advantage of this method is that the operation on the data is free, easy, safe, and accurate. However, the disadvantage is also obvious, since the workload will become overwhelming with increasing data volumes.

Some databases (named derisive databases) are generated from the existing databases (named parent databases). The information of these databases is closely impacted by the databases they are derived from. Since the data deposited in the parent databases are well formatted, automated parsing or generation of new data is possible. A typical derisive database is the protein families database Pfam, which groups the proteins according to the

domains contained. The domain information of Pfam is generated by studying the sequence similarity. The protein sequence is acquired from the protein databases SWISS_PROT and TrEMBL. Databases liked Pfam are not very common, as it acquires information from a sole parent database. Many biological databases have more than one data source, and cannot simply be generated using one program dealing with one data source. Cases will become more complicated while some novel information is hidden in volumes of free text resource. Therefore, multiple programs or integrated programs working with different data sources are desired. The process of the auto data acquirement from various data source is so-called data mining. Due to the fact that the majority of the biological information other than sequence related information comes from the medical literature, the data mining during the database development will focus on the text data mining, which is also known as text mining (TM), or knowledge discovery from text databases (KDT).

Text mining is the process of finding interesting or useful patterns in the corpus of unstructured textual information. TM is not a single algorithm or technique, it is the combination of multidisciplinary techniques such as information retrieval (IR), information extraction (IE), natural language processing (NLP), statistics, and many techniques of data mining, e.g. rule association, decision tree, neural network, genetic algorithm, support vector machines. Though there are different techniques are applied in the TM, generally, the framework of TM can be outlined as below [Dixon *et al.*, 1997]:

1. Information retrieval (IR): The first step is to locate and retrieve relevant documents at hand.

2. Information extraction (IE): The second step is to extract information from the selected documents retrieved in last step.
3. Information mining: This third step is to discover the hidden pattern in the selected documents.
4. Interpretation: This step is the final step to interpret the mined pattern in a natural language.

2.1.3 Procedure of data verification

A good database should present unique and useful information to the users. Above all, the information must be of good quality. The quality of data contains at least two elements: correctness and completeness. Correctness means that the information should be verified by experiment, clinical research, medical report, or some other recognized resource. Any data without verification will lower the credibility of the database. On the other hand, the data should be complete; incomplete information may lead to ambiguity. For example, muscarinic receptor subtypes distributed in different organs and have different functions. Incomplete statement of subtype may cause the information to be ambiguous even wrong. Therefore, quality control of data is critical for the construction and maintenance of a good database. However, it is difficult to automatically perform quality control of the databases, especially for biological databases. The difficulty is mainly due to: different input sources, incomplete information, ambiguous description, redundant information and sometimes typographical error. At present, there is no a standardized nomenclature system available for new protein identification. The designation of new proteins is normally based on the preference of the researchers. Because of the limited information, a protein may have

different names of different proteins may share the same name. It will hardly ever be possible to completely exclude human involvement from biological data annotation. Quality control is not limited to annotation or semantic checks, but also involves the relationships to other parts of the database [George *et al.*, 1987]. Therefore, keeping the information complete, updated and consistent is very helpful to maintain the good quality of data. However, the complexity, incompleteness, and quickly increasing amount of the biological data make it difficult to keep the data updated and in good quality. One possible solution is to withhold suspect data until they have been corrected and supplement the incomplete information when more information is available. Redundancy is another problem often faced when constructing biological databases. It is a consequence of parallel acquisition of the same or highly similar data from independent sources. It also comes from the incompleteness of the information. The redundancy increases the burden of the database searching, and leads to ambiguity of data. To avoid the redundancy, it is necessary to validate the annotation following by similar data searching before the data is added to the database.

2.2 Database Construction

2.2.1 Advantages and classification of database management systems

Database construction is a problem of how to deposit the data in the database and organize them in a specific way. The problem is related to later database representation, database searching and information retrieving. These tasks cannot be served by the database itself, but with the help of a database management system (DBMS). DBMS has advantages in

data independence and efficient access, reduced application development time, data integrity and security, convenient data administration, concurrent access and crash recovery [Ramakrishnan *et al.*, 2002].

Nowadays companies have developed different database management systems to maintain data deposition and application. Generally, there are four types of database systems in the market: the *Flat File Database System*, the *Relational Database Management System* (RDBMS), *Objected Database Management System* (ODBMS) and *Object Relational Database Management Systems* (ORDBMS). Flat files are simply files with a table of information which may be separated by delimiters such as commas, colons, or semi-colons. It was adopted in the early years of the database with limited functions and is not often used any more. RDBMS consists of one or several related tables of simple data in rows and columns. The rows correspond to records, while the columns correspond to attributes or fields in the records. The relationships in the databases are implied by values in specific fields, foreign keys in one table that match those of records in another table. Sometimes the intermediate table containing just the relationships is necessary. RDBMS mainly uses Structured Query Language (SQL) for data definition, data management, data access, and retrieval. ODBMS was developed to deal with complex relationship and store objects. ODBMS stores data in objects rather than in tables. They store attributes and class information, but sometimes they also store methods (behavior) in the database. Other than SQL, ODBMS uses an Object Query Language (OQL) as a standard language for communication. Data in ODBMS are controlled by using the Application Programming Interface (API) or both the API and OQL. Object relational or extended relational, database management systems is the newly emerging database technology in recent years,

which try to unify aspects of both the relational and object databases. ORDBMS are relational databases with data stored in tables, but they have a front end that converts objects to data and data to objects, making it seem to the application that objects are being stored. The comparing of these three modern DBMSs is shown in Table 2.1. Generally, regardless of the difference between DBMSs, database servers include both a server program that serves remote clients and manages the database. They may use some means of standard communication between client and server to allow management of the data such as SQL, OQL, etc. Nowadays, different commercial DBMSs have been developed by companies in these three database types. RDBMS is still the most popular database systems because of simply operation and high efficiency; however, it is believed that ORDBMS servers may someday overtake the RDBMS servers. Table 2.2 lists some representative vendors and their products.

Table 2.1. A Comparison of Database Management Systems

Criteria	RDBMS	ORDBMS	ODBMS
Defining standard	SQL2 (ANSI X3H2)	SQL3/4 (in process)	ODMG-V2.0
Support for object-oriented programming	Poor; programmers spend 25% of coding time mapping the program object to the database	Limited mostly to new data types	Direct and extensive
Simplicity of use	Table structures easy to understand; many end-user tools available	Same as RDBMS, with some confusing extensions	OK for programmers; some SQL access for end users
Simplicity of development	Provides independence of data from application, good for simple relationships	Provides independence of data from application, good for simple relationships	Objects are a natural way to model; can accommodate a wide variety of types and relationships
Extensibility and content	None	Limited mostly to new data types	Can handle arbitrary complexity; users can write methods and on any structure
Complex data relationships	Difficult to model	Difficult to model	Can handle arbitrary complexity; users can write methods and on any structure
Performance versus interoperability	Level of safety varies with vendor, must be traded off; achieving both requires extensive testing	Level of safety varies with vendor, must be traded off; achieving both requires extensive testing	Level of safety varies with vendor; most ODBMSs allow programmers to extend DBMS functionality by defining new classes
Distribution, replication, and federated databases	Extensive	Extensive	Varies with vendor; a few provide extensive support
Product maturity	Very mature	Immature; extensions are new, are still being defined, and are relatively unproven	Relatively mature
Legacy people and the universality of SQL	Extensive supply of tools and trained developers	Can take advantages of RDBMS tools and developers	SQL accommodated, but intended for object-oriented programmers.
Software ecosystems	Provided by major RDBMS companies	Provided by major RDBMS companies	ODBMS vendors beginning to emulate RDBMS vendors, but none offers large markets to other ISVs
Vendor viability	Expected for the major established RDBMS vendors	Expected for the major RDBMS vendors; UniSQL is struggling	Less of an issue than it was; some shakeout still expected
Source: International Data Corporation, 1997			

Table 2.2. Database Management System Products by Vendor

Vendor	**RDBMS**	**ORDBMS**	**ODBMS**
Oracle	Oracle 7.x	Oracle 8.x above	
Sybase	System 10/11		
Informix	Dynamic Server	Universal Server (Illustra)	
IBM	DB/2	Universal Database (DB/2 Extenders)	
UniSQL		UniSQL/X	
Unisys		OSMOS	
Computer Associates	OpenIngres		Jasmine
Gemstone			Gemstone
O2			O2
Object Design			Object Store
Objectivity			Objectivity/DB
Versant			Versant ODBMS
Source: International Data Corporation, 1997			

2.2.2 Consideration of data models for database construction

In addition to the database management system, the data model also play an important role in the database development. Data modeling is the problem of how data are stored in tables and how different data are connected to each other. Three data models are popular in database construction: flat file design, relational design, and object-oriented design as shown in Figure 2.2. The flat file design is the way that all records are normally stored in only one table where each row corresponds to one record and each column corresponds on one attribute (or field) as shown in Table A of Figure 2.2. Since all the annotations are linked as one entity, the access or search of any single field is actually performed on the whole entry. The advantages are easy preparation, construction, and update of database. However, these advantages do no longer exist when the size, dimensions (the number of distinctive fields/information), and relationships between different fields increase. Generally, flat file design is only suitable for small, low-dimensional databases.

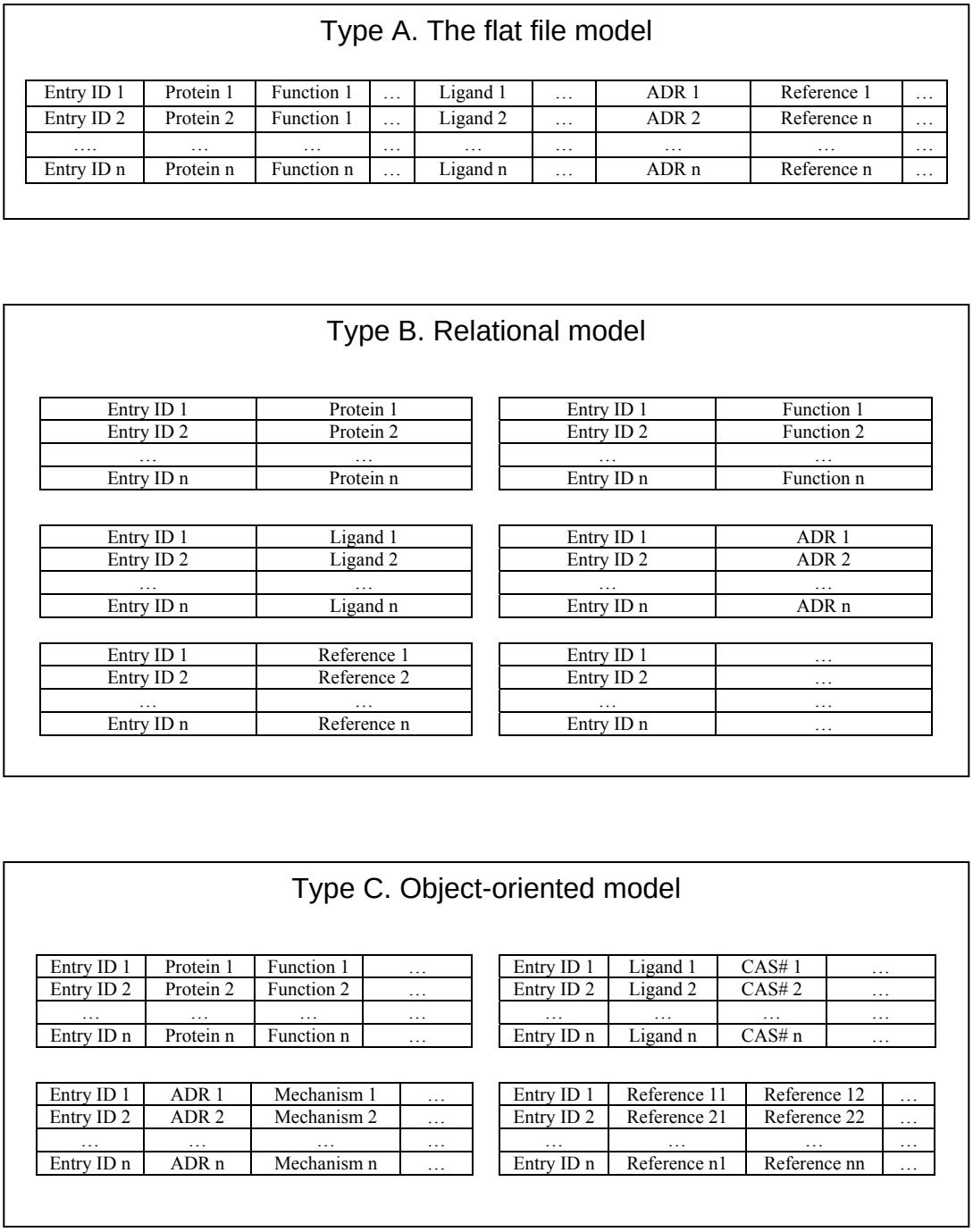
To solve the high-dimensional and complicated data management problem, relational design is introduced. The typical relational model is designed: different tables are created to store distinct information, and each table normally contains limited fields (columns) as shown in Table B of Figure 2.2. These tables are normally linked by unique keys. The uniqueness of the keys helps to solve the complicated relationship between different information fields. For example, it is difficult to build the relationship between protein function and protein domain directly, however, they are both linked with the protein name. Thus introduction of a unique primary (or secondary) key of protein ID, the relationship between function and domain will be easily set up without error. To solve the high-

dimensional problem and prompt searching in and between table(s), indexes are created to access certain field(s) of table(s). Combined with some implemented modules offered by management systems, the representation of data will be more convenient, fast, and flexible. There are many other advantages of relational model, especially when the databases increase in size, dimensionality, and interrelationship. There are also some points, which are difficult to solve by relational data modeling. For example, to solve the high dimensionality and interrelationship problem, many tables are created, which increases the workload of database initialization, upgrade, and evolution. Moreover, any operation on the tables may lead to the regeneration of the new indexes.

Relational data models are well elaborated and widely used in academia and industries. They also show good performance in biological database construction. However, in many cases, it is helpful and necessary to describe data in terms of relationships of groups of records, entries or subsets. For example, the attributes of proteins such as their names, synonyms, gene names, physiological functions, sequence, or structure are always close-bound and presented together, thus there it is not necessary to separate them into different tables. Therefore, an alternative database design, based on object-oriented principles, is introduced to improve the performance of RDBMS. It emphasizes the tight coupling between data and the set of valid operations on that data. Using this design, the database can be dissected into several “functional” components, for example, the protein or ligand prosperities, adverse drug reaction information and reference (Type C, Figure 2.2). The reconstruction of the data into components may slightly lower the search speed, but at the same time, it increases the speed of data retrieval. Most of the biological databases are not so large that the delay of search does not significantly affect the database performance. It

does not mean that object-oriented models are better than relational design or flat file design. The design of the data model is based on the data itself and the consideration of database update, database presentation, database searching, and data retrieval. A good database design should balance of all these factors.

Figure 2.2. The different data model of database



2.3 Database Representation

Database representation is a difficult problem since it is always dependent on the curator's favorite and the solutions vary. Therefore, it is impossible to cover all the technologies in this thesis and only general ideas will be discussed. Considering the security and efficiency of the database, normally, the users don not access the database directly. Instead, access is provided through some applications added to the database. The connection between database servers and the users can be the *Application Programming Interface* (API) or the *Common Gateway Interface* (CGI). For the local access of database, the users send the request to the interface programs; the programs then interpret the request to the database servers; the DBMS will respond to the request and send the result back to the interface programs; finally, the users get their answer from the interface programs. The process is illustrated in Figure 2.3. Compared to the local access, the remote access (normally, the internet access) of database is more popular but complicated. Basically, the whole process of remote access is same as local access, except that the internet factor is included. The rough process of remote access of database is illustrated in Figure 2.4. Users send their requests through the web interface using a form or index; the requests normally are first parsed and packaged to the string suitable for internet transfer, then "posted" or "put" to the HTTP servers, where the requests are further interpreted; the interpreted requests are sent from the HTTP servers to the database servers for information searching and retrieval; the responds of the databases servers are feedback to the HTTP servers and sent back to the users' terminals. The whole process of transferring data among users, HTTP servers and database servers is based on the internet protocol such as HTTP, TCP-IP. Sometimes, the HTTP server, and the database server share the same

machine, however, for security reasons, it is suggested to install these two servers on different machines. The interface plays an important role in connecting the users with the HTTP servers. A user-friendly interface always increases database efficiency. The most popular interfaces include a searchable form or selection list, which is coded using the script languages such as *PERL* script, *JAVA* script, *VB* script, *DELPHI*, and *ASP*. Through the form users can send their requests to the database servers and retrieve the data if available.

Figure 2.3. Illustrate of local database access

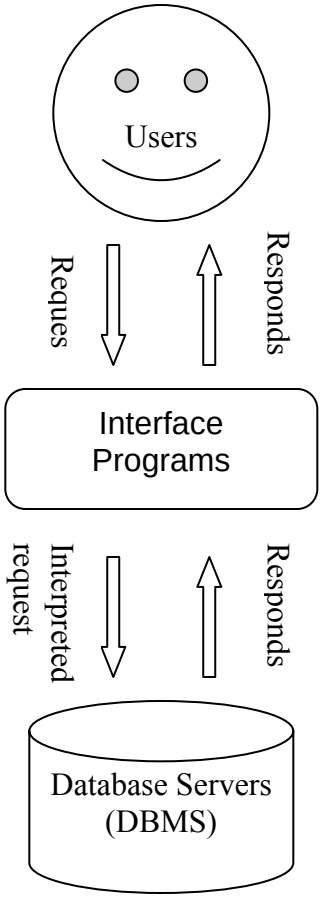
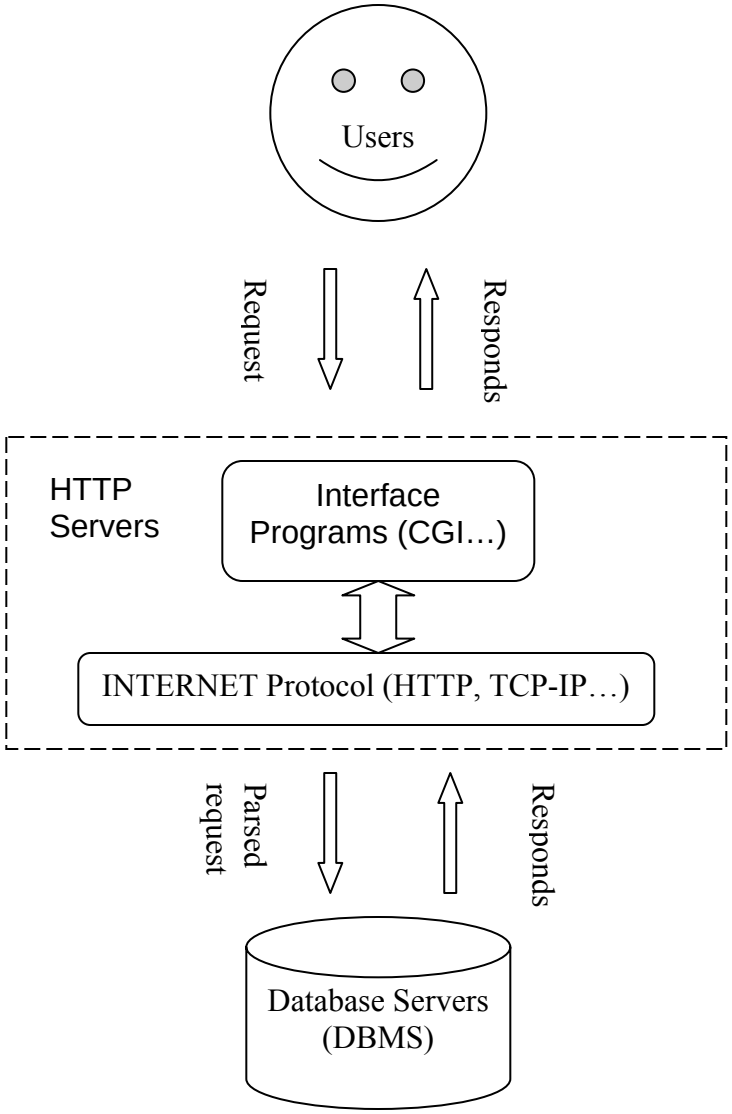


Figure 2.4. Remote access of database



CHAPTER 3 DEVELOPMENT OF TOXICITY DATABASE DART AND ITS APPLICATION IN KNOWLEDGE DISCOVERY

In previous chapter, we briefly introduced the strategy of database development. In this chapter and next chapter we will practically construct two biological databases following this strategy. Furthermore, applications based on these two databases will also carried out to facilitate both biomedical research and drug discovery.

3.1 Development of Drug Adverse Reaction Database (DART)

3.1.1 Collection of ADR targets related information

The majority of drug adverse reaction targets information is collected from the published medical and biological journal papers [McEntyre *et al.*, 2001]. The information also comes from drug review reports submitted by clinical doctors or published in pharmacology textbooks and other publicly accessible literature resources [Pumfor *et al.*, 1997; Wallace *et al.*, 2000; Park *et al.*, 2000; Rang *et al.*, 1999; Klaassen *et al.*, 2001; Baynes *et al.*, 1999]. A target is included in DART if it has been reported that an adverse effect results from direct disturbance of the normal function of the target. The searching of ADR targets from the vast pools of medical literature resource is difficult and time-consuming. To efficiently collect useful and updated information of ADR targets, the latest medical abstract packages, which are freely downloaded from PubMed literature database [McEntyre *et al.*, 2001], are chosen as the major resource. The mining of ADR targets related information is based on the keywords mining under certain rules. The useful keywords are included in two separate keyword lists. One of them contains the

keywords concerning the adverse effects and their synonyms (Table 3.1), which are manually collected from the available medical literatures. The other is the distinct list of protein names along with their synonyms and gene names generated from SWISS_PROT protein database [Bairoch *et al.*, 2000]. An engine is coded using *PERL* scripts to screen the local abstract packages. The abstracts containing ADRs targets related information are selected based on the following rules:

1. The abstract will be picked up for further analysis, if it contains both a keyword or more from the list of the adverse effects and a keyword or more from the list of proteins. This rule will ensure that the abstracts contain both information of adverse effects and proteins.
2. The distance between these two keywords (any combination of two keywords from different keywords lists) should be small enough. This rule is intended to check whether the coexistence of the keywords from adverse effect list and protein list in the abstract are related. The constraint is based on the assumption that two keywords are related if they appear in one complete sentence or adjacent sentences. The assumption is based on experience and the value of distance between two keywords is adjustable.

The rules can be explained by formula as given below:

keywords for adverse effects $A1 = \{w_1, w_2, ..., w_m\}$, (3.1)

keywords for proteins $A2 = \{w_1', w_2', ..., w_m'\}$, (3.2)

collection of abstracts $T = \{t_1, t_2, ..., t_n\}$, (3,3)

and $t(w)$ denotes abstract *t* contains keyword *w*.

if $t_i(w_{j,}, w_k^{'})$ is true, and

$$\min \| w_j \| - | w_k^{'} \| \leq co - related\_dis \tan ce \text{ is true,} \quad (3.4)$$

where $|w|$ is the absolute position in the abstract and co-related_distance is the average maximum distance between two co-related keywords; then the abstract t_i will be picked out for manual analysis.

The abstracts successfully selected from the abstract pools are analyzed manually, and the useful information of ADR targets is collected. Since the mining process is preliminary and manual supervision is heavily involved, evaluation of the system by accuracy is unnecessary. To complement the target information, the full-length papers, reports, and textbooks are also studied with the leads from the abstracts.

Table 3.1. Example of the keywords for text mining

Keyword list
Amnesia
Anaemia
Anxiety
Cardiotoxicity
Confusion
Cytotoxicity
Depression
Edema
Facial flushing
Fever
Gastrointestinal disorders
Headache
Heart related adverse effects
Hepatic adverse effects
Hepatotoxicity
Hyperlipidaemia
Hypertension
Intestinal toxicity
Lipid metabolism adverse effects
Lipodystrophy
Mitochondrial related adverse effects
Muscle related adverse effects
Myopathy
Nausea
Nephrotoxicity
Neurotoxicity
Peripheral neuropathy
Renal failure
Sedation
Sinustachycardia
Vascular adverse effects
Vomiting
Weakness

The key information extracted from literatures is the ADR targets. The targets are associated with diverse adverse effects, and the mechanisms leading to adverse effects are different. According to the mechanisms and the results of adverse effects induced by the drugs, the targets are mainly classified into type A ADRs-related targets, and type B ADRs-related targets respectively:

Type A ADRs-related targets: Targets of this type are responsible for both the adverse effects and their therapeutic effects by interacting with the drugs. In other words, the adverse effects are primarily induced by affecting exactly the same targets as the therapeutic responses achieved by the drugs. Adverse effects induced by the targets of this type include all of the type A ADRs, parts of the type C, and type D ADRs as introduced in Chapter 1. The main mechanisms of the adverse effects in this category are over dose and drug-drug interactions. In most cases, adverse effects of this class are reversible and can be reduced by lowering the drug dose or, in some cases, by changing to a different drug combination. Although the interaction of any drug with its main therapeutic target can potentially induce type A ADRs, only those with well-characterized adverse effects are included in our database at this stage.

Type B ADRs-related targets: Some adverse effects involve the interaction with molecules other than the expected therapeutic targets of the drugs. These molecules (either proteins or nucleic acids) may be the therapeutic targets of other drugs or key molecules in particular biological pathways. Adverse reactions in this category are unpredictable using the dosage or drug-drug interaction analysis. The

adverse effects cover all of the type B, part of type C and type D ADRs. The mechanisms include dysregulation of gene expression, dysregulation of ongoing cell activity, impairment of internal cellular maintenance and impairment of external cellular maintenance [Klaassen *et al.*, 2001]. Compared to the type A-related targets, type B targets should receive more attention, since many of type B adverse effects are unpredictable, serious, and even fatal. In this database, as many as possible type B related targets are collected.

The definition of type A and type B ADR-related targets are not mutually exclusive. There are overlaps between them. Binding of different drugs to a target could induce either type A ADRs or type B ADRs. For example, Beta-1 adrenergic receptor may lead to dysrhythmias when activated by Beta agonists; and cardiac failure may happen while the receptors are inhibited by antagonists. In the database the type of ADR resulting from the binding of a drug to an ADR target is provided. The letter A or B in the bracket before each drug name indicates whether its binding to the corresponding ADR target induces type A or type B ADRs.

In addition to ADR targets, literature-described proteins involved in adverse effect of a chemical are also included in our database and marked as potential ADR targets. Although these potential targets are not yet qualified as ADR targets, they are included in the database based on the expectation that they can potentially become ADR targets for new drugs structurally and chemically similar to the adverse effect inducing chemicals directed at the same target.

Other information in DART includes the molecular properities of each target such as its synonym, name of corresponding gene, physiological function, tissue distribution, and sub-cellular location as well as the adverse effect resulting from the binding of a drug or a chemical to the target. For each ADR target, the type of ADR induced by the binding of a drug is also given. To give a comprehensive perspective about adverse effect targets, more information is provided including the diseases for which each target may play an important role, agonists/antagonists/activators/inhibitors that bind to each target, the related biological pathway (enzyme only) and possible mechanisms of related adverse effect. Cross-links to other databases are also introduced to facilitate the access of information about the function, sequence, 3D structure, nomenclature, and related toxicity literatures of each target.

3.1.2 Data structure and access of database DART

As mentioned above, in total there are 20 distinct information components provided by the Drug Adverse Reaction Target (DART) database. Each of them is allocated in one field of tables in the database. The names of the fields of the concepts, their data types, and their sizes are listed in Table 3.2. The information components and different entries are not unrelated to each other. They are connected through some common items. The relationship map of the components and entries is shown in Map 3.1. To satisfactorily solve the relationship between components and entries, the *Relational Database Management System* (RDBMS), *Oracle 9i* database system, is chosen as our support system. Four tables are generated in the Oracle database: DART_TYPE, LIGAND_TYPE, DART_DATA, and LIGAND_DATA. DART_TYPE assigns each concept field (distinct

information) with one unique type ID as shown in Table 3.2. The unique type ID enables fast searching of respective components and easy update of the data. The detailed information is not stored in table DART_TYPE, but DART_DATA. The DART_DATA table contains three fields: the entry access number (AC), the field type ID, and the information. An example of such table is shown in Table 3.3. The detailed information of all ligands including drugs is stored in table LIGAND_TYPE and LIGAND_DATA. Similar to DART tables, LIGAND_TYPE defines a unique field ID for each concept of ligand information, and the LIGAND_DATA stores the detailed information for ligands associated with their field IDs. Examples of ligand tables are shown in Table 3.4 and Table 3.5. The data are properly formatted before they are loaded into the Oracle database. For the safety consideration, the data tables are not used for web access directly. The readable only reports are thus created in Oracle for information retrieval. To prompt the search process indexes are also created when necessary.

Map 3.1. Relationship of concepts and entries in DART database.

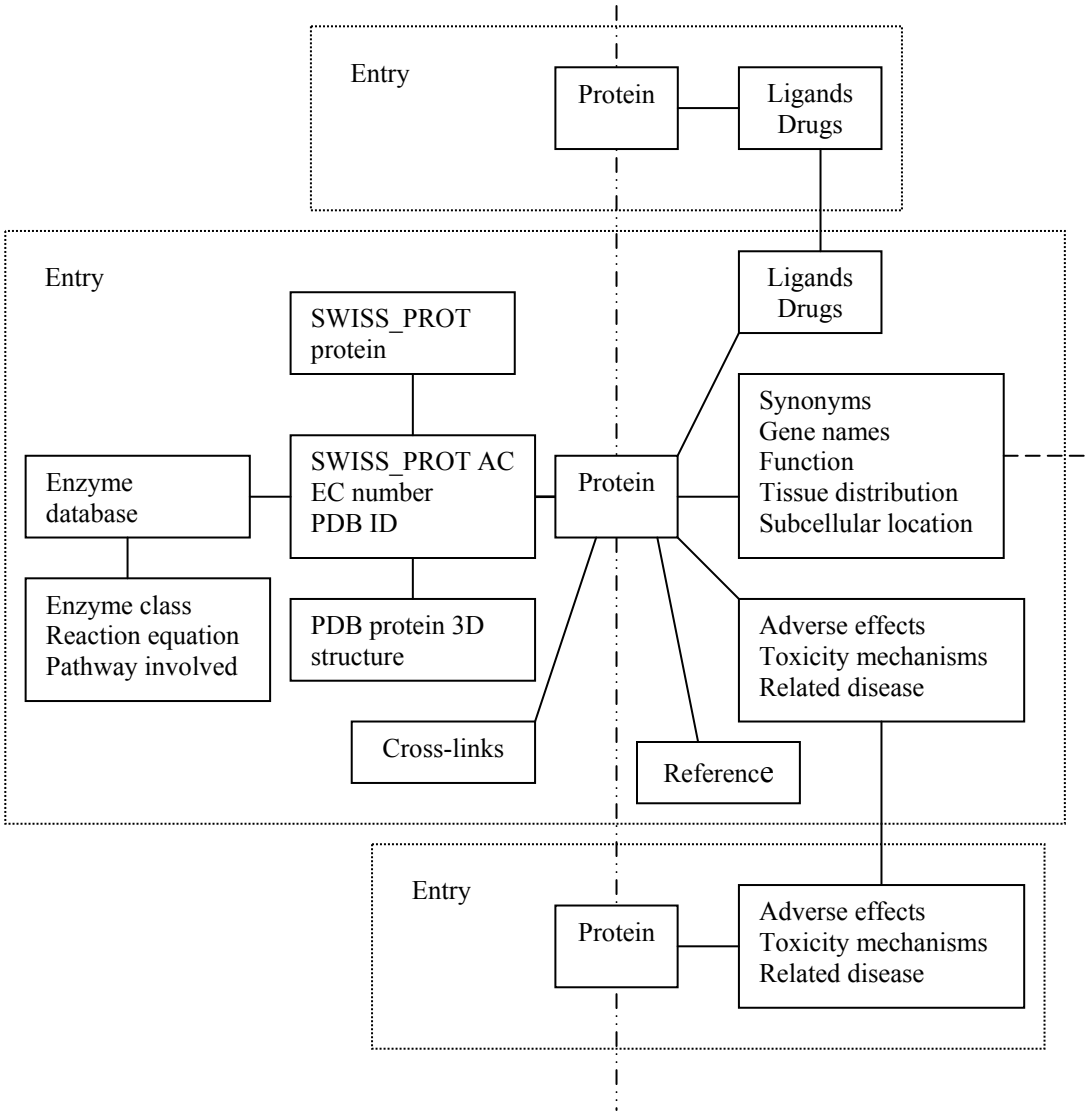


Table 3.2. Data types and sizes of drug adverse reaction targets

ID	Concept	Data Type	Size (bytes)	Occurrence in an entry
801	Protein Name	Text	100	Once
802	Protein Synonym	Text	500	Null, once or more
803	Gene Name	Text	50	Null, once or more
804	EC Number	Text	20	Optional
805	Enzyme Class	Text	20	Optional
806	SWISS_PROT AC	Text	50	Once or more
807	PDB ID	Text	200	Optional
808	Tissue Distribution	Text	200	Optional
809	Subcellular Location	Text	50	Optional
810	Protein Function	Text	200	Once
811	Reaction Equation	Text	50	Optional
812	Pathway Involved	Text	50	Optional
813	Adverse Effect	Text	400	Once or more
814	Related disease	Text	100	Optional
815	Drug (Activator)	Text	500	Optional
816	Agonists / Activators	Text	1000	Once or more
817	Drug (Inhibitor)	Text	500	Optional
818	Antagonists / Inhibitors	Text	1000	Once or more
819	Reference	Text	2000	Once or more
820	Toxicity mechanism	Text	50	Optional

Table 3.3. The examples of Table DART_DATA

AC	ID	VALUE
5	820	Impairment of internal cellular maintenance
…	…	….
6	801	Inducible nitric oxide synthase
6	802	iNOS; NOS, type II; Inducible NOS; Hepatocyte NOS; HEP-NOS
6	803	NOS2A; NOS2
6	804	1.14.13.39
6	805	Oxidoreductases
6	806	P35228
6	807	1NSI 2NSI 4NOS
6	808	Expressed in the liver, retina, bone cells and airway epithelial cells of the lung.
…	…	…
218	819	<<1>> Gregus Z, Klaassen CD. Mechanisms Of Toxicity. Chapter 3, Casarett And Doull¡¯s Toxicology, The Basic Science Of Poisons, 5th Edition.| <<2>> Zhou W, Arrabit C, Choe S, Slesinger PA. (2001) Mechanism Underlying Bupivacaine Inhibition Of G Protein-Gated Inwardly Rectifying K+ Channels. Proc Natl Acad Sci USA;98(11):6482-7
218	820	Impairment of internal cellular maintenance
…	…	…
219	801	Glutamate [Nmda] Receptor Subunit Zata 1

Table 3.4. The data structure of DART ligands

ID	Concept	Data Type	Size (bytes)	Occurrence in an entry
901	Ligand Name	Text	100	Once
902	Ligand Synonym	Text	500	Once or more
903	CAS number	Text	50	Once or more
904	Formula	Text	20	Once
905	Class	Text	20	Optional

Table 3.5. The examples of Table LIGAND_DATA

AC	ID	VALUE
7	905	Anti-urolithic, Antiretroviral, Drug / Therapeutic Agent, Radiation-protective agents
…	…	….
20	901	Benzocaine
20	902	Anesthone; Anesthesin; Anbesol; Americaine; Aethoform; 4-Aminobenzoic acid ethyl ester; Benzocaine; anaesthesinum; ethyl aminobenzoate; ¦thoforme; Ethyl 4-Aminobenzoate; ethylis aminobenzoas; benzocaina; aethylis aminobenzoas; Cepacaine; Parathesin; Orthesin; Auralgan Otic
20	903	94-09-7
20	904	C9H11NO2
20	905	Agricultural Chemical, Anesthetics, Skin / Eye Irritant
…	…	…
118	901	Bleomycin
118	902	bleomycin complex a2/b2; Bleomycin A2; Bleocin; Bleo; bleomycin hydrochloride; Blenoxane; bleomicina; bleomycin sulfate; bleomycin sulfate; N1-(3-(dimethylsulfonio)propyl)bleomycinamide) (Bleomycin A2); BLM; Bleomycin
…	…	…
210	901	Rogletimide

A user-friendly web interface is designed for the remote access of DART database using *ASP* technologies, which is shown in Figure 3.6. This database is searchable by target name or ligand name. It can also easily be accessed by keyword full text search such as adverse reaction, physiological function, or biological pathways. An exact search method through EC access number of enzymes, and SWISS_PROT access number is available as well. Searches involving any combination of these search or selection fields are also supported. The full text search is case insensitive and wild cards are supported. In a query, a user can specify a full name or any part of the name in a text field. Wild characters of '?' and '*' are allowed in text field. Here, '?' represents any one character and '*' represents a string of characters of any length. For example, input of ' cholinesterase ' in the target name field finds entries containing 'cholinesterase' in their name, such as Cholinesterase or Acetylcholinesterase. On the other hand, input of A*cholinesterase finds all the cholinesterase that start their names with 'A'. In this case, '*' represents 'cetyl'.

The result of a typical search is illustrated in Figure 3.7. In this interface, all the toxicity targets that satisfy the search criteria are listed by their protein names and gene names. By clicking the target name, a result table specific to selected target is shown (Figure 3.8), where more comprehensive information about the target is given. The information may include: name of the target, its synonym, gene name, tissue distribution and subcellular location, if known, known agonists/activators/antagonists/inhibitors, possible adverse effect, possible toxicity mechanisms, and literature reference. For an enzyme target, its EC number (Enzyme Data Bank access number, http://www.expasy.org/enzyme/), corresponding biological pathways, and catalytic reaction described in chemical equation form are also provided. Additional information about a target is listed in the target

property item which contains the cross-links to relevant databases such as: SWISS_PROT database [Bairoch *et al.*, 2000], from which the target sequence can be retrieved; The available 3D structure of a target can be accessed through cross-linking to the Protein Data Bank (PDB) database [Berman *et al.*, 2000]. For an enzymatic target, its nomenclature can be obtained from cross-link to the Enzyme Data Bank. The related literature references can be accessed from cross-link to the relevant entries in the PubMed database [McEntyre *et al.*, 2001] and TOXNET (http://toxnet.nlm.nih.gov). More detailed information on ligands, e.g., their synonyms, CAS register numbers, molecular formula and their possible function classification, is also provided, when available, by clicking the ligand name, as shown in Figure 3.9.

Figure 3.6 The search interface of DART.

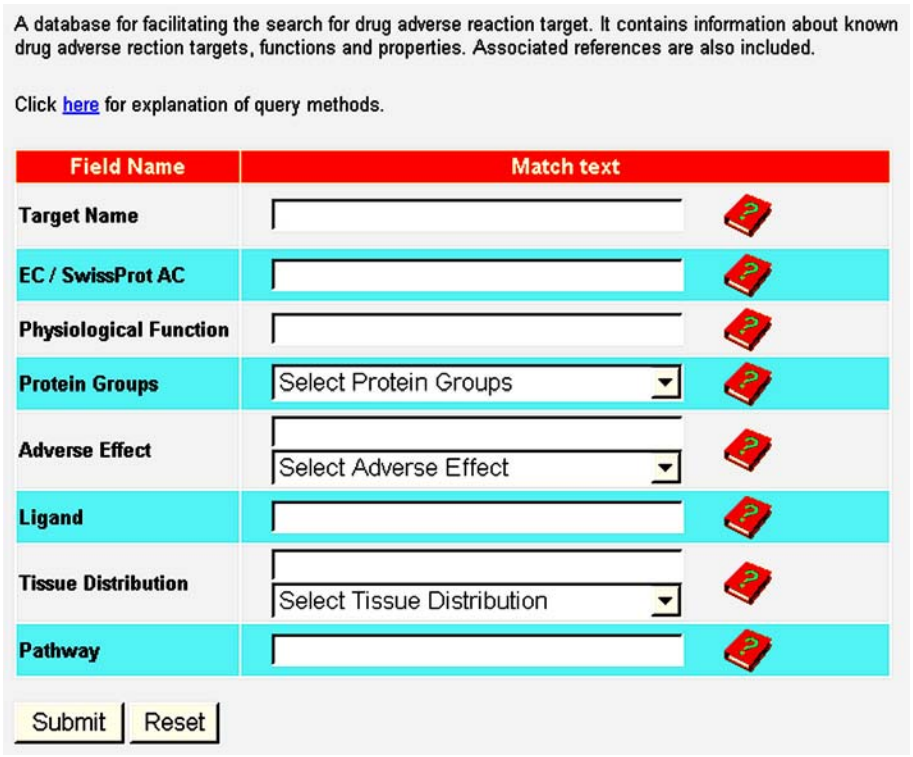
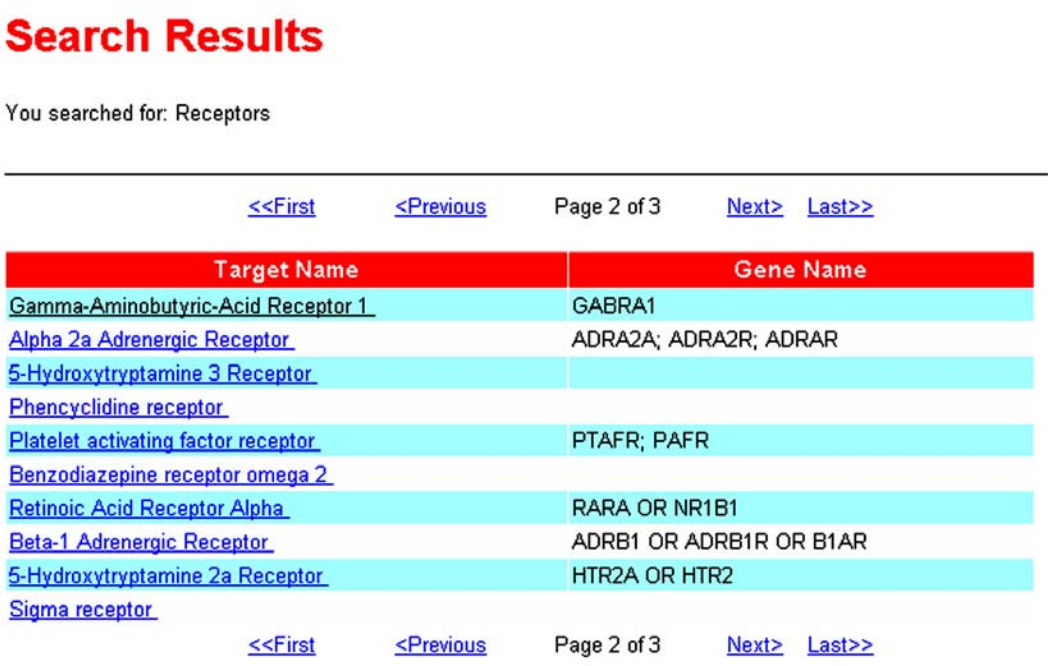


Figure 3.7 The typical search result of DART.



Search Results

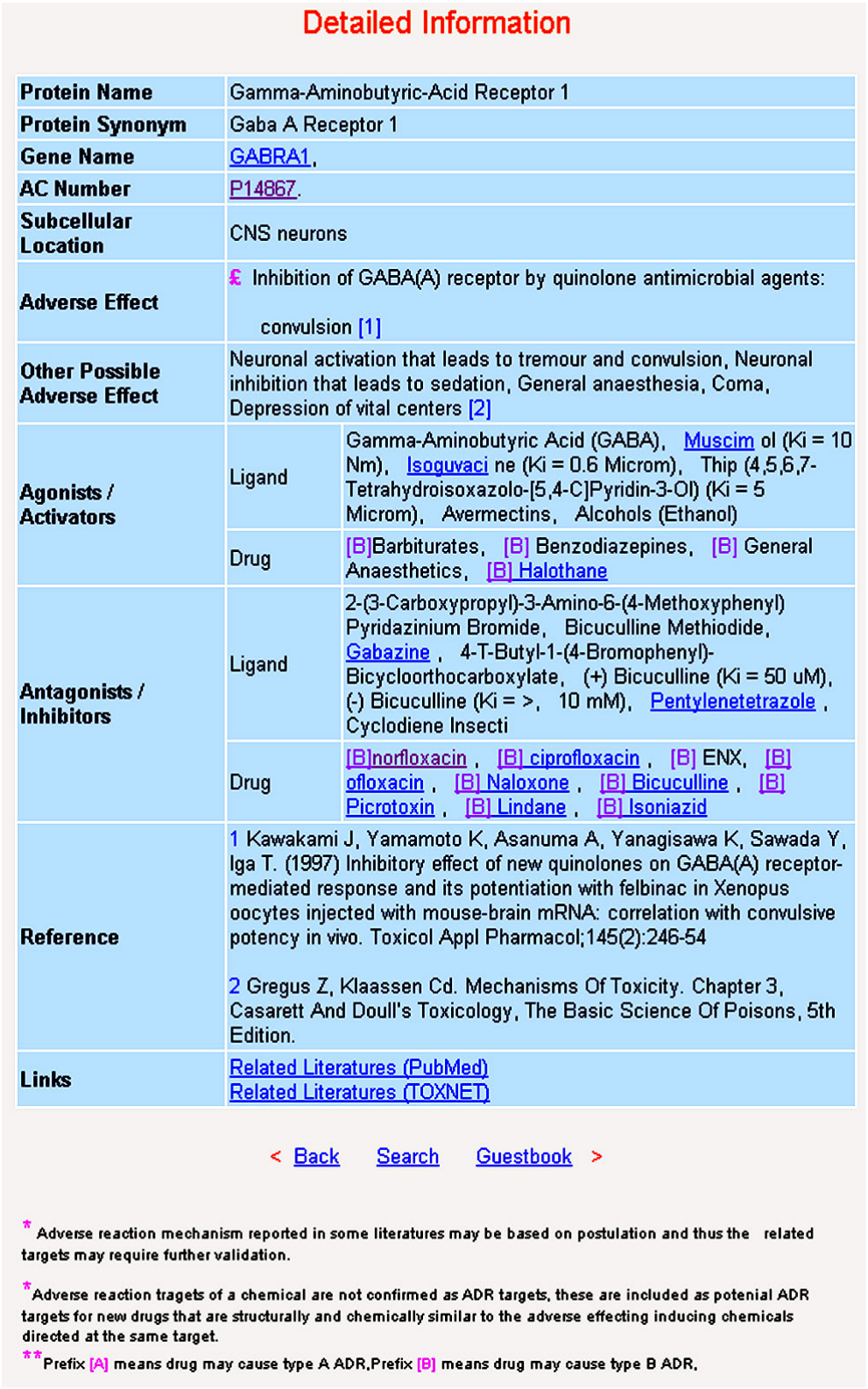
You searched for: Receptors

<<First <Previous Page 2 of 3 Next> Last>>

Target Name	Gene Name
Gamma-Aminobutyric-Acid Receptor 1	GABRA1
Alpha 2a Adrenergic Receptor	ADRA2A; ADRA2R; ADRAR
5-Hydroxytryptamine 3 Receptor	
Phencyclidine receptor	
Platelet activating factor receptor	PTAFR; PAFR
Benzodiazepine receptor omega 2	
Retinoic Acid Receptor Alpha	RARA OR NR1B1
Beta-1 Adrenergic Receptor	ADRB1 OR ADRB1R OR B1AR
5-Hydroxytryptamine 2a Receptor	HTR2A OR HTR2
Sigma receptor	

<<First <Previous Page 2 of 3 Next> Last>>

Figure 3.8. The detailed information of selected toxicity target.



Detailed Information

Protein Name	Gamma-Aminobutyric-Acid Receptor 1	
Protein Synonym	Gaba A Receptor 1	
Gene Name	GABRA1.	
AC Number	P14867.	
Subcellular Location	CNS neurons	
Adverse Effect	£ Inhibition of GABA(A) receptor by quinolone antimicrobial agents: convulsion [1]	
Other Possible Adverse Effect	Neuronal activation that leads to tremour and convulsion, Neuronal inhibition that leads to sedation, General anaesthesia, Coma, Depression of vital centers [2]	
Agonists / Activators	Ligand	Gamma-Aminobutyric Acid (GABA), Muscim ol (Ki = 10 Nm), Isoguvaci ne (Ki = 0.6 Microm), Thip (4,5,6,7-Tetrahydroisoxazolo-[5,4-C]Pyridin-3-Ol) (Ki = 5 Microm), Avermectins, Alcohols (Ethanol)
	Drug	[B]Barbiturates, [B] Benzodiazepines, [B] General Anaesthetics, [B] Halothane
Antagonists / Inhibitors	Ligand	2-(3-Carboxypropyl)-3-Amino-6-(4-Methoxyphenyl) Pyridazinium Bromide, Bicuculline Methiodide, Gabazine , 4-T-Butyl-1-(4-Bromophenyl)-Bicycloorthocarboxylate, (+) Bicuculline (Ki = 50 uM), (-) Bicuculline (Ki = >, 10 mM), Pentylenetetrazole , Cyclodiene Insecti
	Drug	[B]norfloxacin , [B] ciprofloxacin , [B] ENX, [B] ofloxacin , [B] Naloxone , [B] Bicuculline , [B] Picrotoxin , [B] Lindane , [B] Isoniazid
Reference	1 Kawakami J, Yamamoto K, Asanuma A, Yanagisawa K, Sawada Y, Iga T. (1997) Inhibitory effect of new quinolones on GABA(A) receptor-mediated response and its potentiation with felbinac in Xenopus oocytes injected with mouse-brain mRNA: correlation with convulsive potency in vivo. Toxicol Appl Pharmacol;145(2):246-54 2 Gregus Z, Klaassen Cd. Mechanisms Of Toxicity. Chapter 3, Casarett And Doull's Toxicology, The Basic Science Of Poisons, 5th Edition.	
Links	Related Literatures (PubMed) Related Literatures (TOXNET)	

< Back Search Guestbook >

* Adverse reaction mechanism reported in some literatures may be based on postulation and thus the related targets may require further validation.

* Adverse reaction tragets of a chemical are not confirmed as ADR targets, these are included as potenial ADR targets for new drugs that are structurally and chemically similar to the adverse effecting inducing chemicals directed at the same target.

** Prefix [A] means drug may cause type A ADR,Prefix [B] means drug may cause type B ADR.

Figure 3.9. The detailed information of ligand.

Ligand Information

Name	norfloxacin
Synonym	1,4-Dihydro-1-ethyl-6-fluoro-4-oxo-7-(1-piperazinyl)-3-quinolinecarboxylic acid 1-Ethyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-3-quinolinecarboxylic acid 5-23-03-00135 AM-715 BRN 0567897 Baccidal CCRIS 6302 Chibroxin EINECS 274-614-4 MK-366 Norfloxacin Norfloxacine Norfloxacino Norfloxacinum Noroxin 3-Quinolinecarboxylic acid, 1,4-dihydro-1-ethyl-6-fluoro-4-oxo-7-(1-piperazinyl)- 3-Quinolinecarboxylic acid, 1-ethyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-
CAS	70458-96-7
Formula	C16H18FN3O3
Class	Anti-infective agents Antibacterial Drug / Therapeutic Agent Enzyme inhibitors

3.1.3 Statistics and analysis of DART

At present DART database contains a total of 147 confirmed ADR protein targets reported in the literature. The targets distribution with respect to biochemical classes is shown in Figure 3.10 and Table 3.6. Enzymes form the largest group with 74 members, and this is followed by 37 receptors, 19 transporters and 17 other proteins. These 17 other proteins are structural proteins . Further analysis revealed that six classes of enzymes were found in this group including 27 hydrolases, 4 isomerases, 2 ligases, 27 oxidoreductases, 9 transferases, and 5 lyases. These enzymes are distributed in 24 of 61 enzyme sub-families, among which subfamilie of oxidoreductases (with EC1.14.-.-), acting on paired donors with incorporation or reduction of molecular oxygen, shows high probability (17 out of 74) to be the ADR targets. The high concentration of ADR targets in the EC1.14.-.- sub-family is due to the presence of Cytochrome P450 family in ADRs. Cytochrome P450 (CYP) enzymes often play a dominant role in target tissue metabolic activation of xenobiotic compounds. They may also determine drug efficacy and influence the tissue burden of foreign chemicals or bioavailability of therapeutic agents. The competition binding or inhibition of P450s will cause the aberrant metabolism of drugs. Incomplete metabolism of drugs typically leads to the deposition of drugs or their metabolites in organisms resulting in the toxicity effects. Similar phenomenon is observed in the receptor group. As shown in Table 3.6, 20 out of 37 receptors belong to the G-protein coupled receptor family. G protein-coupled receptors (GPCRs) constitute the largest single class of receptors and are responsible for mediating much of intracellular mammalian signaling. Indeed, over 40 percent of the pharmacological targets being evaluated by drug companies are related to G protein-coupled receptors [Benovic *et al.*, 1999]. Thus, any improper

interaction between GPCRs and drugs may cause inappropriate performance of physiological functions resulting in toxic effects. It is noted that, of all these ADR targets, 36 proteins are also found to be therapeutic targets, which likely leads to type A adverse drug effects when their therapeutic performance is significantly affected.

Figure 3.10 The distribution of drug adverse reaction targets

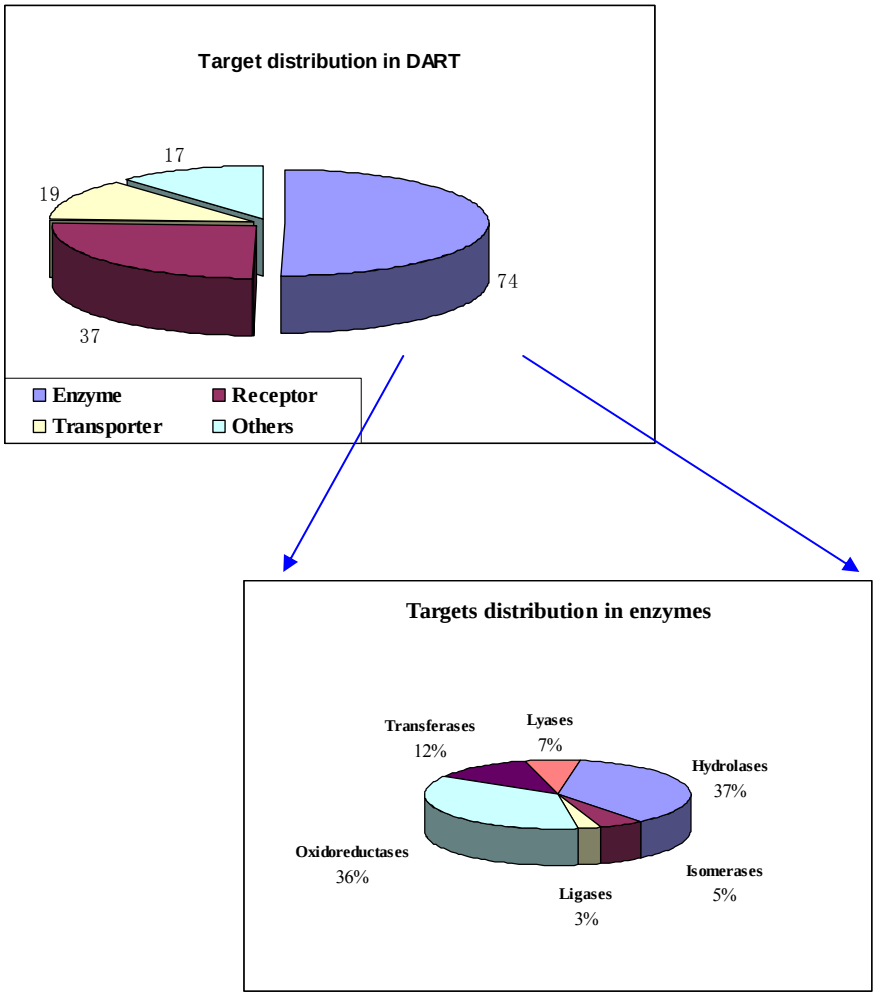


Table 3.6. The distribution of ADR target proteins

Proteins	Family	Definition	Quantity	Total
Enzyme	EC 1.1	Oxidoreductases, acting on the CH-OH group of donors	2	74
	EC 1.10	Oxidoreductases, acting on diphenols and related substances as donors	1	
	EC 1.14	Oxidoreductases, acting on paired donors, with incorporation or reduction of molecular oxygen	17	
	EC 1.3	Oxidoreductases, acting on the CH-CH group of donors	1	
	EC 1.4	Oxidoreductases, acting on the CH-NH2 group of donors	2	
	EC 1.5	Oxidoreductases, acting on the CH-NH group of donors	1	
	EC 1.6	Oxidoreductases, acting on NADH or NADPH	3	
	EC 1.9	Oxidoreductases, acting on a heme group of donors	1	
	EC 2.1	Transferases, transferring One-Carbon Groups	1	
	EC 2.3	Transferases, acyltransferases	1	
	EC 2.6	Transferases, transferring Nitrogenous Groups	2	
	EC 2.7	Transferases, transferring Phosphorus-Containing Groups	5	
	EC 3.1	Hydrolases, acting on Ester Bonds	9	
	EC 3.2	Hydrolases, glycosylases	5	
	EC 3.3	Hydrolases, acting on Ether Bonds	1	
	EC 3.4	Hydrolases, acting on peptide bonds (Peptidases)	8	
	EC 3.5	Hydrolases, acting on Carbon-Nitrogen Bonds, other than Peptide Bonds	2	
	EC 3.6	Hydrolases, acting on Acid Anhydrides	4	
	EC 4.1	Carbon-Carbon Lyases	1	
	EC 4.2	Carbon-Oxygen Lyases	2	
	EC 5.1	Isomerases, Racemases and Epimerases	1	
	EC 5.2	Cis-trans-Isomerases	1	
	EC 5.99	Other Isomerases	3	
	EC 6.3	Ligases, forming Carbon-Nitrogen Bonds	2	
Receptor		G protein coupled receptor	20	37
		Nuclear receptor	6	
		Others	11	
Transporter			19	19
Others			17	17

3.2 Knowledge Discovery from DART: Prediction of ADR Targets Based on Protein Primary Sequence

3.2.1 The need of computational prediction of ADR targets

There are a total of 147 drug adverse reaction targets collected in DART database at present. These targets cover different classes of proteins including enzymes, receptors, transporters, and other proteins, as analyzed previously. It will be useful for facilitating the mechanistic understanding of ADRs if all of these different ADR targets can be identified. Traditionally, ADR targets are identified through studies of patient response to marketed drugs. The identifying process is unsupervised and mainly based subjective evaluation of patients. The toxicity results from animal experimental are often not transferable to humans. It would be of great advantage if those targets could be identified with limited experiments, or even without any experiment, on animals or patients. Sequence analysis is considered a feasible approach for protein function prediction. It is based on the empirical rule that proteins with homologous sequences will fold into similar 3D structure; proteins with similar structures may perform similar functions. According to this rule proteins with high similarity in sequence alignment predicted with *BLAST* or *FASTA* way have a similar function. However, there are exceptions. For example, Glycolate oxidase and IPP isomerase are homologous in sequence alignment, but belong to different enzyme families and behave different functions. Furthermore, classification of ADR targets concerns not only the functions of the proteins, but also their behavior in the biological pathways. Those information cannot be obtained through sequence alignment, thus sequence alignment alone is not sufficient for the classification of ADR targets. Other possible approaches include the structure-function relationship studies, which are limited by the

availability of 3D structures of proteins; biological pathways studies, which are often incomplete. Therefore, in this study we propose a new approach of using Support Vector Machines (SVMs) algorithm to classify ADR-related proteins.

3.2.2 Procedure of ADR-related prediction using SVM classifier

SVMs are principally a two-class classifier. After proper construction of the model of the SVMs using the known example data, the correct classification will be given for the unknown data. One can ignore the performance of the vectors (here, the protein features) in the respective feature spaces. Detailed description of SVM algorithm can be found in the attached **Appendix A**.

The application of SVMs to the classification of ADR targets starts from the assumption that all proteins can be grouped into two classes: the ADR targets (positive data) and non-ADR targets (negative data). The input is the primary sequences of proteins and the output is the respective classes of the proteins. The protein pools came from the SWISS_PROT database, which were divided into three sets of data: Training data, Testing data, and Independent data. All of these three data sets are composed of both positive data and negative data. The training data is used to construct the model; testing data is to examine the model and optimize it; and the independent data can be used for the model evaluation. The design of the datasets is based on the consideration of both the quantities and the qualities of positive and negative data. The selection of positive data is obvious, which include the complete list of the proteins in the DART database. However, the quantity of 147 ADR target proteins is too limited to be divided for training, testing and evaluation of

SVM model. Considering the same proteins in different species may have the same functions, it will be helpful to expand the positive data set by including target proteins from different species. Another advantage of data expansion is that the training datasets and testing datasets will cover all the known ADR targets for machine learning, thus likely increases the accuracy of the classification. To improve the accuracy of positive prediction and well represent true ADR targets, the incomplete proteins such as fragment or partial sequence, as well as the pre-defined proteins such as the hypothetical proteins, are filtered out of the positive data pools.

The selection of negative samples is more complicated and however critical:

1. *Generate the positive data pools*. Collect all target-related proteins (positive data), which include both the ADR targets in DART and same proteins from different species. Searching the latest version of SWISS_PROT knowledge base, a list of SWISS_PROT access number (AC) for the positive data is generated.
2. *Retrieve the family information of the positive data*. Domain families of positive data are generated by matching SWISS_PROT ACs of positive data in the Protein Families Database Pfam. Any Pfam family containing one of the positive SWISS_PROT AC is considered as positive pfam family.
3. *Generate the negative data pools*. The generation of negative families is facilitated by removal of all positive Pfam families from the Pfam database. This process can be conducted by using the Pfam family ID. The proteins in the remaining Pfam families will constitute the negative data pools.
4. *Generate the negative data*. The negative datasets are randomly picked from the negative data pools following the rules that the seed samples of all the negative

Pfam families must be present. The number of the seeds for each negative Pfam family is determined by the size of the negative datasets, which is normally balanced with the positive datasets. Three sets of negative data are thus created for training, testing and independent evaluation.

The positive and negative datasets of protein sequences are now ready for SVM learning process. Every protein sequence can be represented by specific feature vectors assembled from encoded representations of tabulated residue properties including amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility for each residue in the sequence and so on. The constructed feature vectors of both positive samples and negative samples are then loaded into SVM classification system for training in order to identify features that separate positive and negative samples. The SVMPROT system is coded by Dr. Cai [Cai *et al.*, 2003] for protein family classification. After computing the model for about 24 hours for our study, the event-dependent parameters of SVMPROT system are optimized. The well-normalized learning system is now able to classify a new protein sequence into either the positive group (ADR targets) or the negative group (non-ADR targets).

As in the case of all discriminative methods [Baldi *et al.*, 2000], the performance of SVM classification can be measured by the quantity of true positives (*TP*), true negatives (*TN*), false positives (*FP*), false negatives (*FN*), and the overall accuracy (*Q*) given below:

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

Sometimes the performance of SVM is also measured by precision and recall:

$$recall = \frac{N_{correct}}{N_{response}} \tag{3.6}$$

$$precision = \frac{N_{correct}}{N_{key}} \tag{3.7}$$

Where recall represents the number of correct classification $N_{correct}$ returned by the system as a ratio of the total number of relevant classification (or facts) $N_{response}$. Precision is the ratio of the number of correct answers over the total number of answers N_{key}.

3.2.3 Prediction results for ADR targets based on protein sequence

The results for the ADR targets prediction are given in Table 3.7. All the computed *TP*, *TN*, *FP*, and *FN* for the test sets and independent evaluation sets are given in the table. There are 1013 positive protein entries and 2521 negative protein entries used to construct the classification model. Among the 683 positive proteins used for testing model, only 3 of them are incorrectly classified, and there is no incorrect classification out of 2426 negative data during the optimization of model. In an independent evaluation, about 96.8% (735 out of 759) of positive protein sequences are correctly classified into the ADR targets, while 87.3% (1993 out of 2281) of negative proteins are correctly classified into the class of non-ADR targets. The average overall prediction accuracy *Q* is 89.7% for 3040 independent protein entries.

To determine whether or not the prediction error is caused by sequence-related problems such as fragmentation, incomplete chains, or mutations, amino acid sequences of wrongly

predicted proteins are examined. These sequence-related problems do NOT appear to be a significant factor, as they only compose less than 10% of false positive predictions. It is also found that no significant (less than 5%) sequence-related problems exist in false negative predictions. Several other factors may affect the prediction accuracy. One is the uncertainty of the protein definition in the SWISS_PROT. Notably about 68.8% of false positive predicted proteins are hypothetical proteins. Those proteins still need further verification and study for their physiological functions. On the contrast, hypothetical proteins were not included in the positive samples. The different treatment of the positive and negative data set is under the consideration for protecting the feature completeness of positive data and diversity of negative data. Another possible reason is the diversity of protein samples, which could be the main reason for the false negative prediction. It is likely that not all possible types of proteins are adequately represented in training and testing datasets. This can also be improved along with the availability of more protein data.

To further verify the accuracy of the computational prediction of ADR targets, randomly selected proteins are used for SVMPROT prediction. It is satisfying that the overall prediction accuracy Q is as high as 93.3%. The partial predicted results are listed in Table 3.8. The physiological functions of proteins are provided to help understanding of their possible roles in the normal physiological activities and etiology. The prediction is also supported by possible (indirect or incomplete) experimental/clinical results. More research is needed to fully confirm our prediction.

Table 3.7. Results of ADR target proteins predicted by SVMPROT

Training set		Testing set*				Independent evaluation set				
Positive	Negative	Positive		Negative		Positive		Negative		***Q***
		TP	*FN*	*TN*	*FP*	*TP*	*FN*	*TN*	*FP*	%
1013	2521	680	3	2426	0	735	24	1993	288	89.7

Q: overall accuracy;
TP: true positives
FN: false negatives
TN: true negatives
FP: false positives

* The testing results is demonstrated with
Recall = 99.6%
Precision = 100%

Table 3.8. Comparing physiological functions of predicted ADR targets with their possible adverse effects

Protein	Philological Function	Possible Adverse Effect	Predicted Result
Cytochrome P450 1A2	Involved in an NADPH-dependent electron transport pathway in liver microsomes. Known to oxidize a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics.	Tardive dyskinesia	Positive
P-Glycoprotein 1	Responsible for decreased drug accumulation in multidrug-resistant cells. Protection of intestine mucosa layer. Energy-dependent efflux pump.	Inhibition causes lower oral bio-availability of drug to targeted tissues	Positive
Androgen Receptor	Involved in the regulation of eukaryotic gene expression. Found to affect cellular proliferation and differentiation in target tissues.	Cytotoxicity, Cell death	Positive
Cytochrome P450 2D6	Responsible for the metabolism of many drugs and environmental chemicals that it oxidizes. It is involved in the metabolism of drugs such as antiarhythmics, adrenoceptor antagonists, and tricyclic antidepressants.	Arrhythmias, Bradycardia, Confusion, Opioid dependance, Lactic acidosis, Hepatotoxicity	Positive
Pyruvate Kinase	Pyruvate kinase is the final regulatory point in the catabolic Embden-Meyerhoff-Parnas pathway, which controls the carbon flux of glycolytic intermediates and regulates the level of ATP in the cell.	Emolytic anemia	Positive
Tyrosine Aminotransferase	Known to transaminate dicarboxylic as well as aromatic amino acids.	Eye and skin lesion, Neurological problems	Positive
Adenylosuccinate Synthetase	Involved in the de novo pathway of purine nucleotide biosynthesis.	Toxicologic effects of L-alanosine	Positive
Heat Shock-Related 70 Kda Protein 2	Known to stabilize preexistent proteins against aggregation and mediate the folding of newly translated polypeptides in the cytosol as well as within organelles, in cooperation with other chaperones. Found to provide a driving force for protein translocation in mitochondria and the endoplasmic reticulum. Involved in signal transduction pathways in cooperation with hsp90. Known to participate in all these processes through their ability to recognize nonnative conformations of other proteins. Found to bind extended peptide segments with a net hydrophobic character exposed by polypeptides during translation and membrane translocation, or following stress-induced damage.	Cardiotoxicity, Anaemia, Gastrointestinal disorders	Positive
Voltage/Ca2+ Activated K+Channel	Known to play an important role in controlling membrane potential and contractility of urinary bladder smooth muscle (UBSM).	Convulsion, spasm	Negative

3.3 Application of DART: Computational Evaluation of Drug Safety

3.3.1 The need for the development of computer-aided drug safety evaluation tools

Drug discovery process is complex, costly and time-consuming. The process is also known to be risky during the pre-clinical and clinical testing phases due to the potential ADR effects of compounds. In the pre-clinical phase, toxicity experiments in animal and cells (mainly the drug metabolism enzymes P450s) are standard methods for the drug safety evaluation. However, the toxicity results of a drug found in animal experiments do not always agree with its adverse effects in human; and enzyme analysis is focused on the potential adverse effects induced by the drug-drug interaction. In the clinical testing phases, the knowledge of ADRs mainly comes from the clinical reports of healthcare professionals submitted to the ADRs reporting systems. Such systems do provide accurate ADR information of potential drugs; unfortunately, the information is based on individual patient evaluation. Therefore, an efficient and safe method for drug safety evaluation is desired.

Computer-aided assessment of potential ADR of investigational drugs has received increasing attention in recent years. In this work, a specific computer method is explored for identification of the putative ADR-causing target proteins. The approach involves the use of a protein-ligand software and a supporting database. The computer program adopted here is the inverse docking algorithm (INVODOCK) developed by Chen [Chen, 2001], which is able to automatically predict the possibility of docking a ligand into a protein. The supporting database is the target database, the Drug Adverse Reaction Target

database (DART), which collects the literature-confirmed drug adverse reaction target proteins.

3.3.2 A drug safety prediction method: INVODOCK and its algorithm

INVODOCK is a ligand-protein inverse docking algorithm, which is able to facilitate computer-automated inverse docking search for finding putative protein targets of a small molecule [Chen *et al.*, 2001]. The proper application of the software requires a protein cavity database developed from relevant protein entries in PDB. The database contains models of individual cavity in each protein. A cavity model is a cluster of overlapping spheres that fill-up that cavity [Kuntz *et al.*, 1982]. A drug is flexibly docked into the cavity by a procedure involving multiple conformer shape-matching alignment of the molecule to the cavity. This is followed by the molecular-mechanics torsion optimization and energy minimization on both the ligand and the binding region of the protein. Putative protein targets are selected based on a new scoring scheme that performs binding competitive analysis in addition to the evaluation of molecular mechanics ligand-protein interaction energy. More detailed description of the core docking algorithm is provided in following.

Docking is a term used for computational schemes that attempt to find the "best" fit in the binding of two molecules: a protein and a ligand. The molecular docking problem can be defined as follows: *Given the atomic coordinates of two molecules, predict their "correct" bound association*. In figurative way, the docking problem can be described as the relationship between "lock" and "key": how to put the suitable "key" into the suitable

"lock". The early docking algorithms were based mainly on geometric criteria [Kuntz *et al.*, 1982], although a few energy-based algorithms were also developed [Goodsell *et al.*, 1996]. Generally, there are three key ingredients in the docking: (1) representation of the system, (2) conformational space search, and (3) ranking of potential solutions [Halperin, 2002]. Representation of the system is the question of how to define a protein surface. The basic description of the protein (or ligand) surface is the atomic representation of exposed residues. However, this explicit simulation is usually based on real potential energy functions, which require significant computer time for modeling complex binding. Different mathematical models have been built to simplify the system by assuming one or both of the molecules in the docking system are rigid. Considering this approximation, the computational procedure inherent to docking can be classified into three levels: (1) Rigid body docking, a highly simplistic model, which treat the two molecule as rigid solid bodies; (2) Semi-flexible docking, which consider one of the molecule, normally the small ligand, flexible and the other molecule rigid; (3) Flexible docking, which both molecules are regarded as flexible, although the flexibility is usually limited.

The docking procedure induces energy changes and conformational changes. The conformational changes may be due to the motion of backbone or side chain of the molecules, especially at the binding site, solvation, electrostatics, or van der Waals force. Theoretically there exists one conformation of docked complex which has minimum energy:

$$E_{total} = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_\phi}{2}[1 + \cos(n\phi - \gamma)] + \sum_{non-bond} 4 \in \left[\left(\frac{\sigma}{R} \right)^{12} - \left(\frac{\sigma}{R} \right)^{6} \right] - \frac{q_1 q_2}{DR} \tag{3.8}$$

E_{total} is the total energy of the molecule; $\sum_{bonds} K_r(r - r_{eq})^2$ is the energy of bond stretches, where K_r is the force constant of bond stretches, r is the length of bond and r_{eq} is the length of bonds at equilibrium position. $\sum_{angles} K_\theta(\theta - \theta_{eq})^2$ is the energy of angles blending, where K_θ is the force constant of bond angles, θ is the angle of bond bending and θ_{eq} is the angle of bonds at equilibrium position. $\sum_{dihedrals} \frac{V_\phi}{2}[1 + \cos(n\phi - \gamma)]$ is the energy of bond torsion, where V_ϕ is the energy cost associated with the deformation, is the value of the dihedral angle in our structure, n (the multiplicity) is the number of energy minima that exist for the dihedral angle, γ is used to determine where the dihedral angle passes through an energy minimum. $\sum_{non-bond} 4 \in \left[\left(\frac{\sigma}{R}\right)^{12} - \left(\frac{\sigma}{R}\right)^{6}\right]$ is van der Waals interaction energy for pairs of non-bonded atoms, where σ is the distance at which the Lennard-Jones interaction energy is zero, $\in$ is the minimum energy of the Lennard-Jones potential, R is distance between atoms. $\sum_{non-bond} \frac{q_1 q_2}{DR}$ is the charge interaction energy for non-bonded atoms, where q_1 and q_2 are the partial charges on the atoms, D is the dielectric constant.

There are numerous ways for two molecules to interact, and the number of possibilities grows exponentially with the size of components [Shoichet *et al.*, 1992]. Hence, docking is actually the problem of searching conformational space, which will locate the most stable state (global minimum) in the energy landscape in a rugged funnel shape. The solutions can be obtained by two essentially different approaches: (1) full solution space search, which scans the entire solution space in a predefined systematic manner; (2)

gradual guided progression through solution space, which scans only part of the solution space in a partially random and partially criteria-guided manner or generates fitting solution [Halperin, 2002]. The solutions for the conformational space search are assessed by some energy functions in a coarse manner (e.g. hinge-bending algorithms [Sandak *et al.*, 1998]) or more rigorously (e.g. Monte Carlo [McMartin *et al.*, 1997], molecular dynamics [Wang *et al.*, 1999], genetic algorithm [Politowska *et al.*, 2002] etc). Whichever algorithm is chosen, the objective is to find the energy minimum state of docking.

There are different algorithms available for docking that can produce an immense number of solutions for the conformational space search. Thus, a scoring scheme is desired to discriminate between the "correct" native solutions and other low-energy complex conformations. Theoretically, free-energy simulation can be applied, however, it is not practical due to its computational requirements. The alternative solutions of the scoring problems have been succeed by, for example, evaluating the similarity to a reference structure [Fradera *et al.*, 2000]. At present, there is no efficient method available for reliable discrimination between correct solutions and false positives generated by predictive docking algorithms [Norel *et al.*, 1999].

3.3.3 Procedure of identifying potential ADRs targets of 11 marketed anti-HIV drugs

The computational study is carried out for the prediction of potential ADR target of 11 marketed AIDS drugs. These drugs have been used in the treatment of AIDS by binding to two different therapeutic targets. They are of different structural types including HIV

protease inhibitors, nucleoside reverse transcriptase inhibitors, non-nucleoside reverse transcriptase inhibitor and nucleoside analogs [Table 3.9]. The 3D structures of these drugs are downloaded on-line and some of them are shown in Figure 3.11.

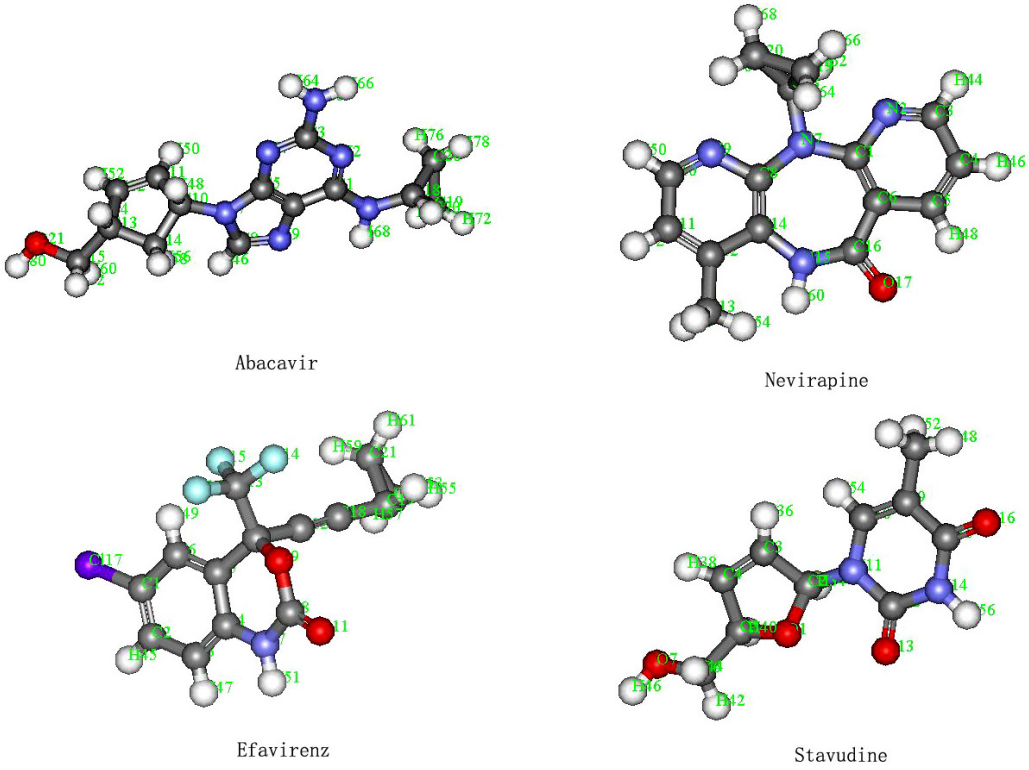
The prediction of ADR targets by INVODOCK is carried out on UNIX machine with the CPU speed of 700MHz and memory of 128Mbytes. The INVODOCK is initialized by transforming the respective 3D structures of ligands (drugs) into ".pdb" format. The species of cavity database (PDB protein database) is selected to constrain the search scope to reduce the computational burden and increase the relevance of the search. In this study, only the mammalian and human PDB proteins are chosen for docking. The selection is based on the objective of studying the possible toxicity targets in human. The inclusion of mammalian proteins is supplements limited human PDB proteins, whose structure are evolutionarily close to each other. INVODOCK docks each ligand to multiple proteins. The time of calculation is varied from one week to two weeks depend on the ligand sizes. The small ligands may take longer time to search all the possible cavities because they are more possible to be docked into protein cavities compared to larger ligands.

Table 3.9. The clinical observed side effects of AIDS drugs

Chemical Name	Brand Name	Therapeutic mechanism*	Observed side-effects
Delavirdine	Rescriptor	NNRTI	Severe rashes and skin reactions are common
Nevirapine	Viramune	NNRTI	Severe rashes and skin reactions are common
Efavirenz	Sustiva	NNRTI	Dizziness, headache, insomnia, inability to concentrate, and rash are common. Vivid dreams, nightmares, and hallucinations can occur.
Lamivudine	Epivir, 3TC	NA	The most common side effects are gastrointestinal: upset stomach, nausea, vomiting. A more serious side effect neutropenia, a blood disorder, can occur. Pancreatitis or inflammation of the pancreas can also occur, but is more common in children.
Stavudine	D4H, Zerit	NRTI	Peripheral neuropathy, numbness, tingling, or pains in hands or feet are common.
Abacavir	Ziagen	NRTI	The common side effects are potential hypersensitivity reaction, fever, stomach and bowel symptoms, general malaise, and sometimes rash. If resume the Abacavir after stopping, the hypersensitivity can occur quicker and more severely.
Amprenavir	Agenerase	PI	The most common adverse effects are nausea, diarrhea, vomiting, and oral rashes. Elevated blood glucose levels and altered body fat distribution are also reported.
Indinavir	Crixivan	PI	It is generally well tolerated, and sometimes may cause mild elevation of bilirubin, and some people may get kidney stones.
Nelfinavir	Viracept	PI	Nelfinavir is generally well tolerated. Diarrhea, nausea, abdominal pain and elevated blood glucose levels can occur sometimes.
Ritonavir	Norvir	PI	Ritonavir can cause nausea, vomiting, diarrhea, a feeling of numbness, altered tastes, and a rise in cholesterol levels, however the side effects can be reduced when drugs are taken with food.
Saquinavir	Invirase	PI	Saquinavir is generally well tolerated. Diarrhea, nausea, and abdominal pain sometimes can occur.

* NNRTI: non-nucleoside reverse transcriptase inhibitor;
NRTI: nucleoside reverse transcriptase inhibitor;
NA: nucleoside analogs;
PI: protease inhibitor.

Figure 3.11. Examples of the 3D structures of the AIDS drugs



3.3.4 Prediction results of anti-HIV drugs and analysis

The program INVODOCK produces a list of docked proteins, file "ligand-docklist" (the "ligand" is the names the respective studying drugs) after all the cavity searching finished. This "ligand-docklist" file lists all the proteins into which the studied ligand can be docked, along with their PDB IDs. These proteins are considered as the potential target proteins of the corresponding ligand. Moreover, the individual docking results for each protein are saved in ".pdb" format as well (Figure 3.12).

The putative targets of investigational drugs predicted by INVODOCK are summarized in Table 3.8, which is listed in quantities of human targets against that of all studied species (mammalians). It is noted that if investigational drug can be docked into more putative targets (both the human targets and total targets), it may induce more dangerous or various the drug side effects. The conclusion is also supported by examination of the number of the predicted ADR targets. The identification of ADR targets is by matching the putative target proteins with proteins in the ADR target database DART. The matched of ADR targets are also listed in Table 3.10. It is found that comparing to the non-nucleoside reverse transcriptase inhibitors (NNRTI) and nucleoside reverse transcriptase inhibitors (NRTI), the HIV protease inhibitors (PI) have few number of ADR targets. This finding agrees with the clinical observation that protease inhibitors have less serious side effects such as mitochondrial toxicity-related side effects, comparing to other types of AIDS drugs. Therefore, PIs are commonly used for the treatment of first-line therapy or naïve therapy since their first use in 1996 [Cvetkovic *et al.*, 2003; Joly *et al.*, 2002].

To confirm that the predicted ADR targets are meaningful and accurate, two comparisons are made. First, whether the side effects induced by the predicted ADR targets agree with the clinical side effects of drugs, which are predicted to be docked into the target proteins. The example of this comparison is shown as Table 3.11. The protein DNA polymerase beta is suggested to be the putative targets of 9 AIDS drugs. Our result indicates that 6 out of 9 drugs show the kind of adverse side effects expected for the binding to DNA polymerase beta (+), one has side effects probably related to this protein (+/-), and the remaining two drugs are not reported to produce effects related to this protein (-). The near 80% prediction accuracy demonstrates that the computational prediction of toxicity by INVDOCK program is feasible. The binding of some AIDS drugs such as Abacavir [Kakuda *et al.*, 2000], Lamivudine [Hart *et al.*, 1992; Kakuda *et al.*, 2000], Stavudine [Stammberger *et al.*, 1989; Kakuda *et al.*, 2000] with DNA polymerase beta has been confirmed by experimental studies.

The efficiency of the prediction is also studied by comparison of the clinical side effects of studied drugs with the side effects induced by corresponding putative ADR targets. As shown in Table 3.12, there are 4 proteins predicted by INVODOCK to be the putative ADR targets of Abacavir, among which DNA polymerase beta contribute the gastrointestinal side effects and the neuropathy such as hypersensitivity, and DNA topoisomerase I may be related to the fever, general malaise, and rash. It is not reported that Abacavir causes kidney related disease or Lipodystrophy, which are induced by FK506-binding protein 1A and Sterol regulatory element binding protein-1 respectively. The putative protein targets basically explain the clinical side effects of the studying drugs, however, some possible side effects induced by the predicted putative ADR targets have

not yet been found clinically. The reason could be either the over-prediction of the program or the inadequate clinical information.

Figure 3.12. The docking result of AIDs drugs into putative protein target.

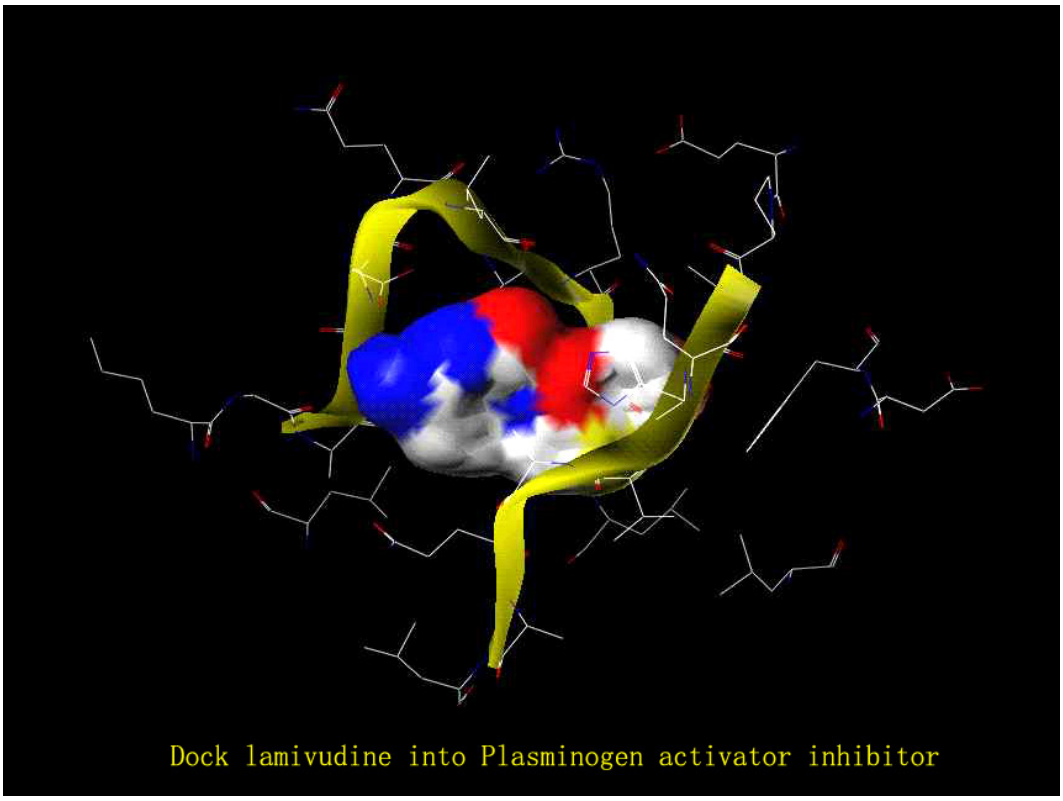


Table 3.10. The targets predicted by INVODOCK

Drug	Target proteins*	ADR targets**	
Abacavir	70/136	4	DNA polymerase beta
			DNA topoisomerase I
			FK506-binding protein 1A
			Sterol regulatory element binding protein-1
Amprenavir	15/30	2	DNA polymerase beta
			DNA topoisomerase I
Delavirdine	22/45	2	DNA polymerase beta
			DNA topoisomerase I
Efavirenz	47/92	4	DNA polymerase beta
			DNA topoisomerase I
			Insulin
			Sterol regulatory element binding protein-1
Indinavir	2/3	1	DNA polymerase beta
Lamivudine	70/182	6	DNA polymerase beta
			DNA topoisomerase I
			FK506-binding protein 1A
			Insulin
			Macrophage migration inhibitory factor
			Sterol regulatory element binding protein-1
Ritonavir	1/1	0	-
Saquinavir	3/6	0	-
Nelfinavir	17/22	1	DNA polymerase beta
Nevirapine	72/131	3	DNA polymerase beta
			DNA topoisomerase I
			Sterol regulatory element binding protein-1
Stavudine	66/125	5	DNA polymerase beta
			DNA topoisomerase I
			Insulin
			Macrophage migration inhibitory factor
			Sterol regulatory element binding protein-1

* Quantities of target proteins predicted by INVODOCK, shown in human proteins/total proteins.
** The INVODOCK-predicted ADR target proteins found in DART database.

Table 3.11. Relationship between the predicted ADR target protein of different AIDs drugs and their corresponding observed side effects

ADR target	Side effect induced	Drug	Observed drug side effect	Correlation*
DNA polymerase beta	pancreatitis, peripheral neuropathy, abdominal pain, adverse GI effects such as nausea and vomiting	Abacavir	Hypersensitivity reaction, fever, stomach and bowel symptoms, general malaise, and rash.	+
		Amprenavir	Nausea, diarrhea, vomiting, oral rashes, elevated blood glucose levels and altered body fat distribution.	+
		Delavirdine	Severe rashes and skin reactions.	-
		Efavirenz	Dizziness, headache, insomnia, inability to concentrate, rash, vivid dreams, nightmares, and hallucinations.	+
		Indinavir	Mild elevation of bilirubin, and kidney stones.	+/-
		Lamivudine	Nausea, diarrhea, vomiting, oral rashes, elevated blood glucose levels and altered body fat distribution.	+
		Nelfinavir	Diarrhea, nausea, abdominal pain and elevated blood glucose levels	+
		Nevirapine	Severe rashes and skin reactions.	-
		Stavudine	Peripheral neuropathy, numbness, tingling, or pain in hands or feet.	+

*The co-relation between the predicted side effect of the ADR target proteins and clinical observed side effect of the drugs:

"+" means correlated;
"-" means non-correlated;
"+/-" means uncertain relationship;

Table 3.12. Relationship between the different predicted ADR targets of the AIDs drug and its corresponding observed side effects

Drugs	Observed Side effect	Name	Protein Side effect	Correlation*
Abacavir	The common side effects are potential hypersensitivity reaction, fever, stomach and bowel symptoms, general malaise, and sometimes rash. If resume the Abacavir after stopping, the hypersensitivity can occur quicker and more severely.	FK506-binding protein 1A	Nephrotoxicity	-
		DNA topoisomerase I	Cytotoxicity and DNA lesion.	+/-
		DNA polymerase beta	Pancreatitis, peripheral neuropathy, abdominal pain, adverse GI effects such as nausea and vomiting.	+
		Sterol regulatory element binding protein-1	Lipodystrophy, increased lipogenesis, and insulin resistance.	-

CHAPTER 4 DEVELOPMENT OF KINETIC DATABASE KDBI AND ITS APPLICATION IN KNOWLEDGE DISCOVERY

4.1 Development of Kinetic Data of Bio-molecular Interactions (KDBI)

4.1.1 Collection of kinetic information

The majority of the information in Kinetic Data of Bio-molecular Interactions database (KDBI) is manually collected from the published medical literature manually. The information includes experimentally determined kinetic data for protein-protein, protein-DNA, protein-RNA, protein-ligand, DNA-ligand, and RNA-ligand interactions. KDBI provides detailed description about binding or reaction events, participating molecules, binding or reaction equations, kinetic data, and related references. A variety of molecular descriptions are also provided which include names of molecules, synonyms, SWISS_PROT AC for a protein or CAS number for a small molecule ligand, molecular formula, classification, protein function and tissue distribution. The kinetic data is presented as one or a combination of kinetic quantities as given in the literature of a particular event. These quantities include association/dissociation rate constant, on/off rate constant, first/second/third/… order rate constant, catalytic rate constant, equilibrium association/dissociation constant, inhibition constant, and binding affinity constant.

4.1.2 Data structure and access of database KDBI

There are 20 distinct information components provided by the database of KDBI. Each component is allocated in one field of table in the database. The data structure of the database such as the fields, their data types, their sizes and their occurrence in one entry is listed in Table 4.1. The distinct information components and different entries are cross-related to each other. For example, one protein can interact with different ligands and one ligand can also interact with different proteins vice versa. The relationship map of the concepts and entries is shown in Map 4.1. To properly organize the complicated relationship between the molecules, the unique molecule ID is assigned to each distinct molecule. Two specific characters are prefixed in the unique molecule ID to differentiate the molecule types, following by the five numbers for different molecules. The identification of molecules is defined as shown in Table 4.2. The properly formatted data is loaded into the supporting database system, *Oracle 9i* database system, which is the *Relational Database Management System* (RDBMS). Four tables are generated in the Oracle database to solve the complicated relationships: KDBI_TYPE, MOL_TYPE, KDBI_DATA, and MOL_DATA. KDBI_TYPE assigns a unique field ID for each of the distinct kinetic information, as shown in Table 4.1. The detailed content of kinetic information is stored in table KDBI_DATA. The KDBI_DATA table contains three fields: the entry access number (AC), the field ID, and the respective kinetic contents. The example of the table is shown at Table 4.3. The detailed information of bio-molecules and their data structure is defined and stored in table MOL_TYPE and MOL_DATA respectively, which MOL_TYPE defines the unique field IDs for each concept of ligands information, and the MOL_DATA stores the detailed information of proteins and ligands associated with their field IDs. The examples of molecule tables are shown in Table 4.4 and Table 4.5. The data are analyzed and formatted before they are loaded into the Oracle

database. For safety consideration, the data tables are not directly accessed by web users. Instead, the readable-only reports are created in Oracle for information retrieval. To enable fast-speed for searching, indexes are also created when necessary.

Map 4.1. Relationship of KDBI contents and entries

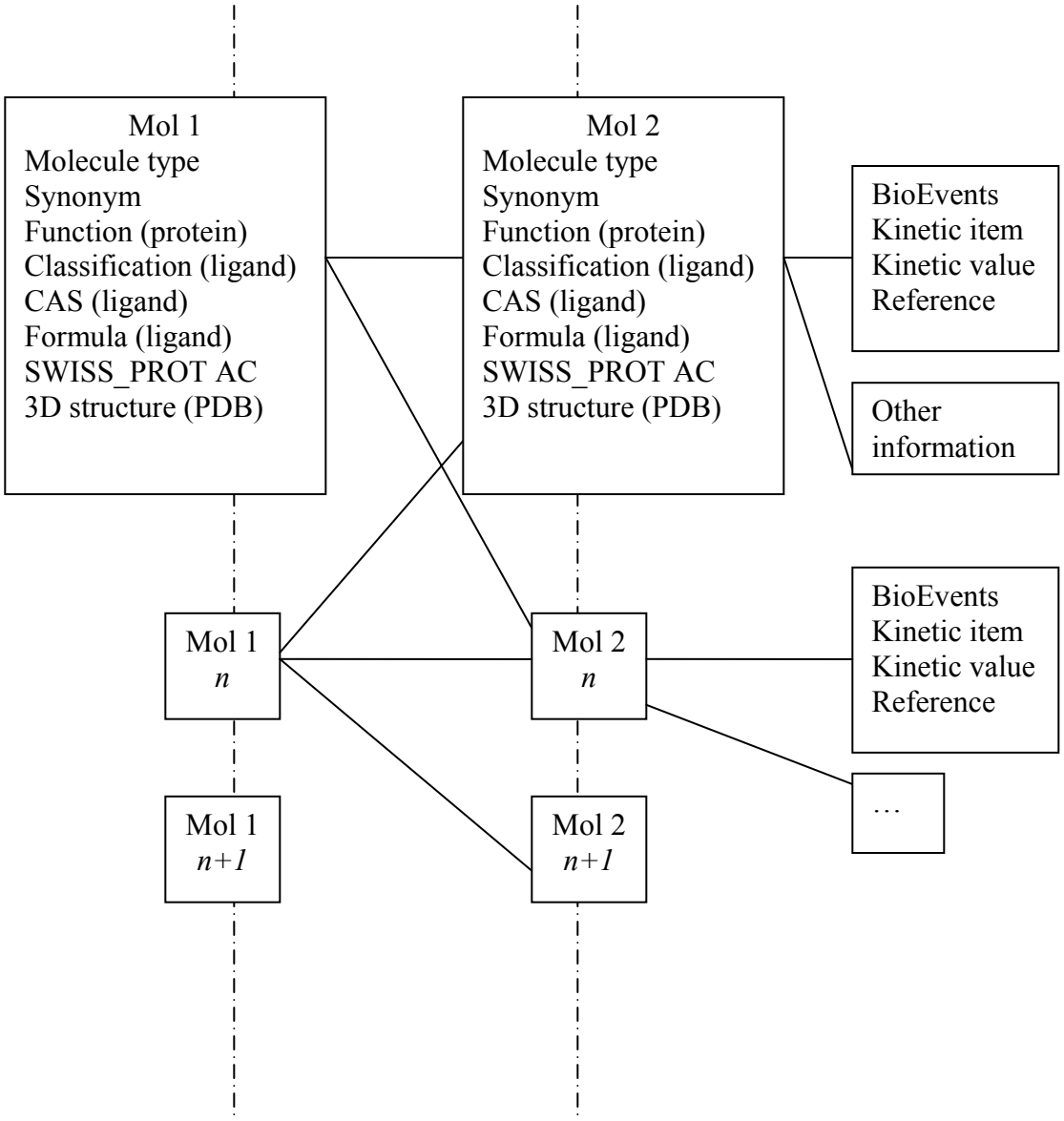


Table 4.1. The data structure of Table KDBI_TYPE

ID	Concept	Data Type	Size (bytes)	Occurrence in an entry
701	Molecule ID one	Text	10	Once
702	Molecule ID two	Text	10	Null or once
703	BioEvent	Text	100	Once
704	Reaction	Text	200	Null, once or more
705	Kinetic Item	Text	20	Once
706	Kinetic value	Text	20	Once
707	Kinetic unit	Text	20	Once
708	Reference	Text	200	Once

Table 4.2. The definition of molecule ID

Prefix	**Molecule type**
PN	Protein
PC	Protein Complex
NA	Nucleic Acids
LG	Ligand
IN	Ion
CL	Cell or membrane

Table 4.3. The example of KDBI_DATA

AC	TYPE	VALUE
…	…	…
5	701	PN00091
…	…	…
21	701	PN00013
21	702	LG00102
21	703	Inactivation of NOS by 2-ethylaminoguanidine
21	704	
21	705	Second order rate constant k2
21	706	6.70E-05
21	707	M-1s-1
21	708	Wolff DJ, Gauld DS, Neulander MJ, Southan G. (1997) Inactivation of nitric oxide synthase by substituted aminoguanidines and aminoisothioureas. J Pharmacol Exp Ther; 283(1): 265-73
…	…	…
510	701	PN00132
510	702	PN00054
510	703	The inhibition of mini-collagenase by TIMP-1
…	…	…

Table 4.4. Data structure of molecule table MOL_TYPE

ID	Concept	Data Type	Size (bytes)	Occurrence in an entry
711	Molecule ID	Text	10	Once
712	Name	Text	20	Once
713	Synonym	Text	200	Null, once or more
714	Type	Text	20	Once
715	CAS RN	Text	100	Once or more (ligand)
716	Formula	Text	50	Once (ligand)
717	Classification	Text	100	Optional
718	SWISS_PROT AC	Text	100	Once or more (protein)
719	Function	Text	200	Once (protein)
720	Tissue Distribution	Text	100	Optional
721	3D Structure Infomation (PDB)	Text	100	Optional
722	Nucleic Acid Information (NDB)	Text	20	Optional

Table 4.5. Examples data of table MOL_DATA

AC	TYPE	VALUE
…	…	…
32	711	PN00007
…	…	…
255	711	PN00013
255	712	Nitric oxide synthase isoforms
225	713	NOS
255	714	Protein
255	715	
255	716	
255	717	
255	718	
255	719	Produces nitric oxide (NO) which is implicated in vascular smooth muscle relaxation through a cgmp-mediated signal transduction pathway. No mediates vascular endothelial growth factor (vegf)-induced angiogenesis in coronary vessels and promotes blood clot.
…	…	…
470	711	LG00007
470	712	Oxygen
470	713	CCRIS 1228| EINECS 231-956-9| HSDB 5054| Hyperoxia| LOX| Liquid oxygen| Oxigeno [Spanish]| Oxygen| Oxygen molecule| Oxygen, liquified| Oxygen-16| Oxygene |Pure oxygen
…	…	…

Database Access

KDBI has a web interface at http://xin.cz3.nus.edu.sg/group/kdbi/kdbi.asp, which is shown in Figure 4.1. The kinetic data for a binding event is searchable by several methods. One method is via the name of participating molecules (protein, nucleic acid, small peptide, ligand or ion). An entry can also be searched through a SWISS_PROT AC number for a protein or the CAS number for a small molecule ligand. Moreover, keyword-based text search is also supported. To facilitate convenient access to relevant data, a partial list of proteins is provided. Searches involving combination of these methods or selection fields are also supported.

The search is case insensitive and wild cards are supported. In a query, a user can specify full name or any part of the name in a text field. Wild character of '*' and '?' is allowed in text field. Here, '?' represents any one character and '*' represents a string of characters of any length. For example, input of 'reductase' in the molecule field finds entries containing 'reductase' in their name, such as NADPH-adrenoferredoxin reductase, NADH-CoQ recuctase, cytochrome P450 reductase, 5-alpha reductase, thiol-disulfide oxidoreductase, etc. On the other hand, input of NAD% reductase finds all the reductase start their names with 'NAD'. In this case, '*' represents ' PH-Adrenoferredoxin' and 'H-CoQ' respectively.

The result of a search is illustrated in Figure 4.2, in which all events that satisfy the search criteria are listed. This list includes the name of the participating molecules as well as the description of the corresponding event. The related kinetic data can be obtained by clicking the “Kinetic data” button of a selected event. The page of kinetic data, as shown in Figure 4.3, provides detailed description about the reaction equation (while available),

the kinetic data given in the literature, and the source of the literature. Further information about the participating molecules can be obtained by clicking the name of the respective molecules. As shown in Figure 4.4, the corresponding molecular information page provides the name, synonym, SWISS_PROT access number for a protein or CAS number for a small molecule ligand, classification and formula for a small molecule ligand (while available), and the function, tissue distribution and cross-link to SWISS_PROT database [Bairoch, 2000] for a protein. Moreover, hyperlinks are provided to facilitate access to the relevant reference in Medline and available 3D structural entries in PDB [Berman *et al.*, 2000]. For a nucleic acid, hyperlink to its available 3D structural entries in NDB is also provided [Berman *et al.*, 1992].

Figure 4.1. Web interface of KDBI.

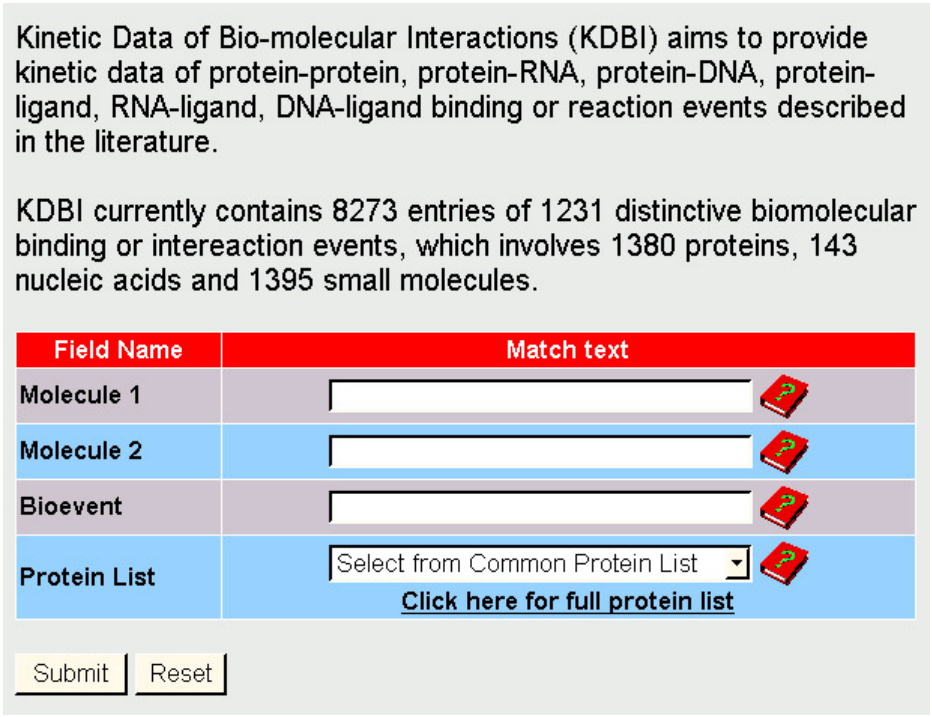


Figure 4.2. The interface for a search in KDBI.

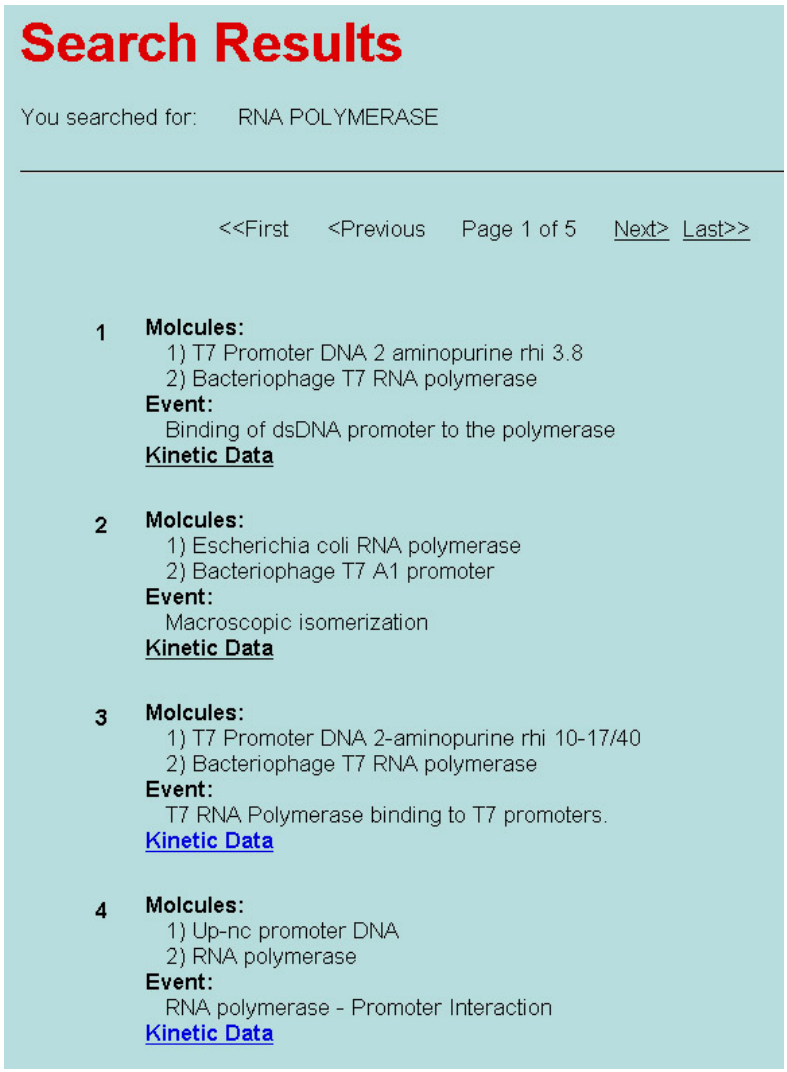
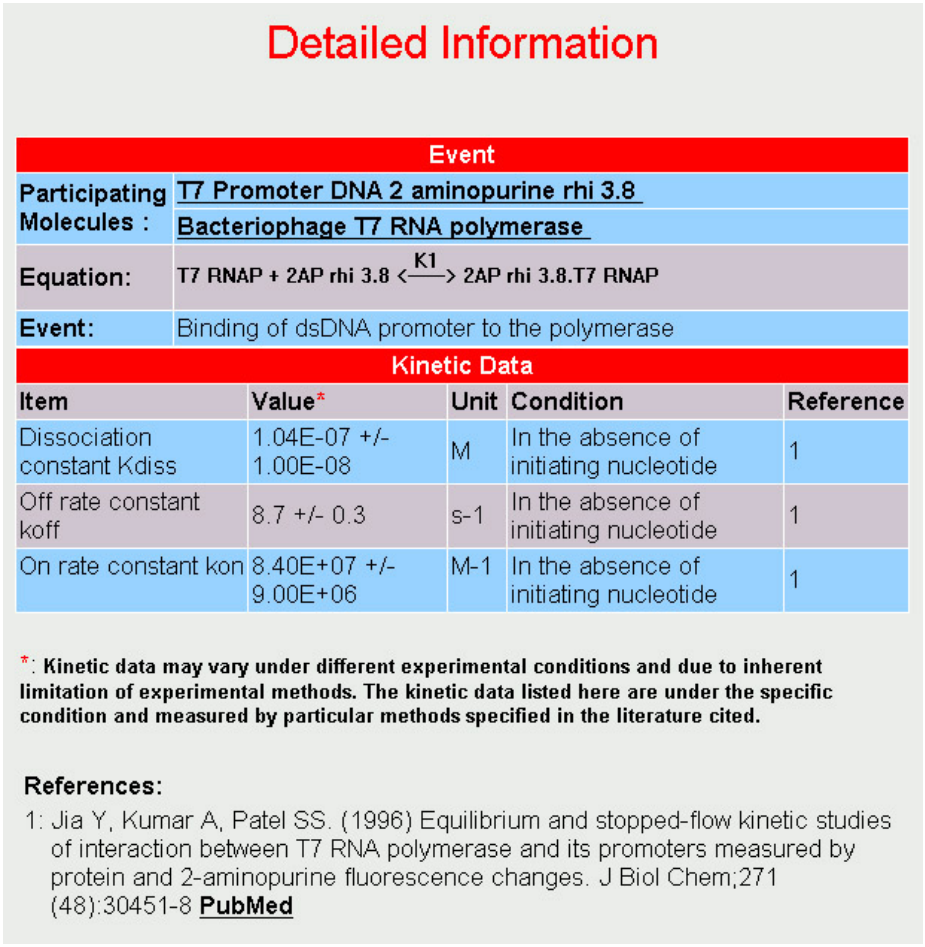


Figure 4.3. The kinetic data page.



Detailed Information

Event	
Participating Molecules :	T7 Promoter DNA 2 aminopurine rhi 3.8
	Bacteriophage T7 RNA polymerase
Equation:	T7 RNAP + 2AP rhi 3.8 $\xleftrightarrow{K1}$ 2AP rhi 3.8.T7 RNAP
Event:	Binding of dsDNA promoter to the polymerase

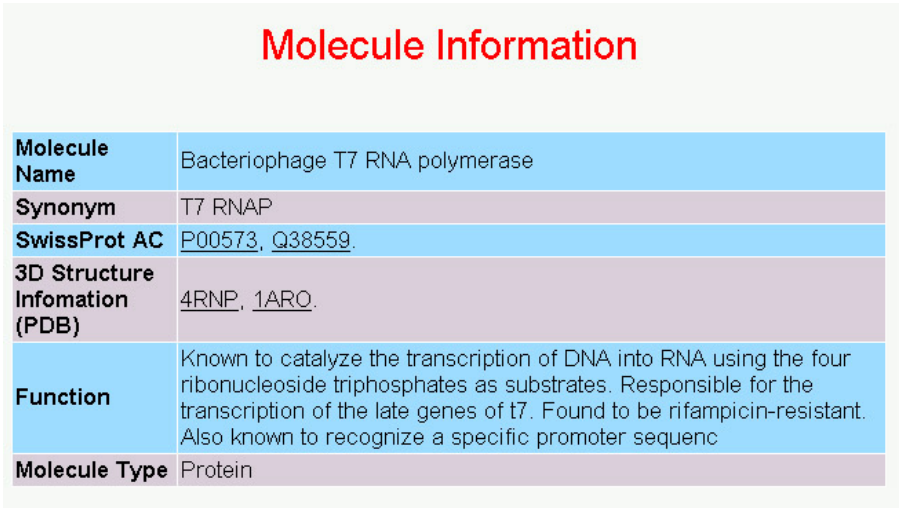
Kinetic Data				
Item	Value*	Unit	Condition	Reference
Dissociation constant Kdiss	1.04E-07 +/- 1.00E-08	M	In the absence of initiating nucleotide	1
Off rate constant koff	8.7 +/- 0.3	s-1	In the absence of initiating nucleotide	1
On rate constant kon	8.40E+07 +/- 9.00E+06	M-1	In the absence of initiating nucleotide	1

*: **Kinetic data may vary under different experimental conditions and due to inherent limitation of experimental methods. The kinetic data listed here are under the specific condition and measured by particular methods specified in the literature cited.**

References:

1: Jia Y, Kumar A, Patel SS. (1996) Equilibrium and stopped-flow kinetic studies of interaction between T7 RNA polymerase and its promoters measured by protein and 2-aminopurine fluorescence changes. J Biol Chem;271 (48):30451-8 **PubMed**

Figure 4.4. Molecular information page.



Molecule Information

Molecule Name	Bacteriophage T7 RNA polymerase
Synonym	T7 RNAP
SwissProt AC	P00573, Q38559.
3D Structure Infomation (PDB)	4RNP, 1ARO.
Function	Known to catalyze the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates. Responsible for the transcription of the late genes of t7. Found to be rifampicin-resistant. Also known to recognize a specific promoter sequenc
Molecule Type	Protein

4.1.3 Statistics and analysis of KDBI

KDBI currently contains 8273 entries of biomolecular binding or reaction events. There are a total of 1380 proteins, 143 nucleic acids, and 1395 small molecules (ligands and ions) included in the database. The distribution of the molecules is shown in Figure 4.5. The majority of the interactions kinetically studied by researchers are ligand-protein, protein-protein, and protein-nucleic acids interactions, which represent 45.7%, 25.4%, and 15.3% of all the bio-molecular interaction entries, respectively (Table 4.6). It is noted that this statistics on the bio-molecular interactions is based on the incomplete collection of published kinetic information from year 2002 to year 1990. Work is still underway to collect kinetic data published in earlier years or in other resources. This is expected to significantly increase the number of entries of the database in the near future. This distribution of bio-molecular interactions containing kinetic information may be changed after new data are added. The kinetic information provided in KDBI is diverse, and includes the rate constants such as association rate constant, dissociation rate constant and order rate constants and different equilibrium constants. The types of kinetic constants collected by KDBI are list in Table 4.7. Studying the entries in detail, several facts are found: 1. Different kinetic analysis methods on the same interaction (same in both molecules and bio-event) may lead to the different kinetic value. However, the difference is normally not significant. 2. For the same interaction measured by same method, the value of same kinetic items may be significantly different due to different experimental condition. Thus, a complete entry should include both the method and condition as well as the kinetic value.

Figure 4.5. The distribution of molecules in KDBI database

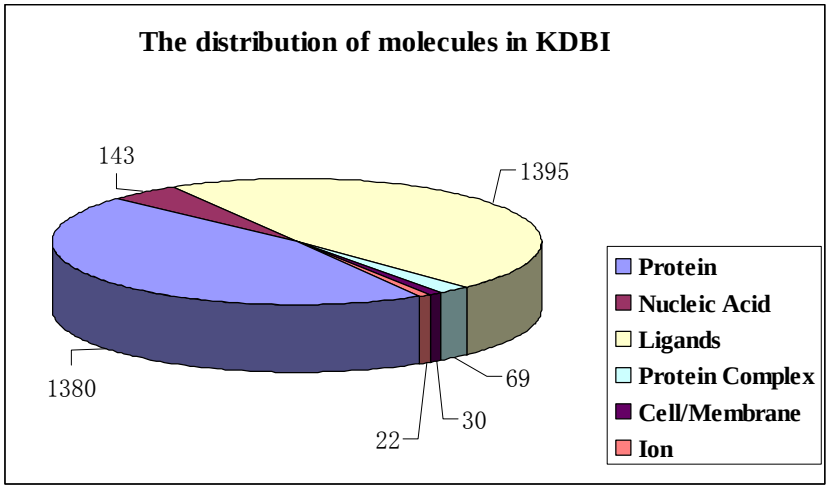


Table 4.6. The distribution of molecular interaction in KDBI

Molecule Interactions	Entries	Percentage of all entries (%)	Distinct Interactions
Ligand - Ligand	195	2.4	78
Ligand - Nucleic Acids	115	1.3	52
Ligand - Protein	3783	45.7	1697
Protein - Protein	2105	25.4	927
Protein - Nucleic Acid	1263	15.3	720
Nucleic Acids - Nucleic Acids	434	5.2	326
Cell – Others	73	0.9	41
Ion – Others	187	2.3	108
Others	121	1.5	76
Total	8276		4025

Table 4.7. The types of kinetic constants information provided by KDBI

Rate constant	**Equilibrium constant**
Association rate constant (k_{ass})	Association constant (K_{ass})
Dissociation rate constant (k_{diss})	Dissociation constant (K_{diss})
Order rate constant (k_1, k_{-1}, k_2…)	Affinity constant (K_a)
Cytalytic rate constant (k_{cat})	Inhibition constant (K_i)
Inhibitary rate constant (k_i)	Michael constant (K_m)
On rate constant (k_{on})	Order constant (K_1, K_{-1}, K_2…)
Off rate constant (k_{off})	
Channel open rate constant (k_o)	
Channel close rate constant (k_c)	
Forward rate constant (k_f)	
Backward rate constant (k_b)	
Observed rate constant (k_{obs})	
Apparent rate constant (k_{app})	
Peudo-order rate constant (k_p)	
Exchange rate constant (k_e)	
Bimolecular rate constant (k_b)	

4.2 Knowledge Discovery for KDBI: Construction of Protein-Protein Interaction Network

4.2.1 The need of the construction of protein-protein interaction network

It is estimated, from the study of human genome, that there are about 100,000 or more proteins in a human body. A portion of them has been well studied for their structures and functions by experiment. The functions of some of the proteins are probed by means of sequence analysis including sequence homology and share a motif with proteins of known function. In many instances, a protein carries out its function by interaction with other molecules and these interactions are essential in a broad spectrum of biological processes including regulation of metabolic pathways, immunologic recognition, DNA replication, progression through the cell cycle, and protein synthesis [Alberts *et al.*, 1989]. The construction of the protein-protein interaction network is thus a useful approach to facilitate the understanding of the mechanisms of various biological processes. There are several protein-protein interaction databases available on-line. The Biomolecular Interaction Network Database (BIND) [Bader *et al.*, 2003] archives biomolecular interaction, complex and pathway information. It provides for the storage of interactions or complexes involving all kinds of biomolecules (protein, DNA, RNA, and small molecules) down to the atomic level and provides a home for annotated research findings and externally published scientific data. The information of protein complex derives from mass spectrometry based proteomics as well as from high-throughput Genetic interactions. The Database of Interacting Proteins (DIP) documents experimentally determined protein-protein interactions, which is useful for understanding protein function and protein-protein relationships, studying the properties of networks of interacting proteins, benchmarking

predictions of protein-protein interactions, and studying the evolution of protein-protein interactions. The BindingDB [Chen *et al.*, 2001] is a public web-accessible database of measured binding affinities for biomolecules, genetically or chemically modified biomolecules, and synthetic compounds. In addition to these databases focused on protein-protein interaction, there are some pathway databases available on-line, which represent the biological processes or protein interactions as block diagrams. These databases include the metabolic pathway databases KEGG [Kanehisa *et al.*, 2002], the Signaling PAthway Database (SPAD) [http://www.grt.kyushu-u.ac.jp/spad/], the Biochemical Pathways of ExPASy [http://www.expasy.org/cgi-bin/search-biochem-index]. These databases are useful tools for facilitating the understanding of the protein functions and biological processes. Some pathway databases also provide the quantitative information such as the Database of Quantitative Cellular Signaling, which includes reaction schemes, concentrations, rate constants, as well as annotations on the pathway models [Sivakumaran, 2003]. Such databases provide further insights into cellular functions and processes; however, the kinetic studies are limited to the known pathways. Since quantitative understanding of biological processes should explicitly specify the kinetics of all chemical reaction steps in a pathway, it is very meaningful to construct a protein-protein interaction network with kinetic information. Here, in this work, such protein-protein interaction network is constructed for the kinetic study and visualization of protein-protein interaction. As shown in the Table 4.6, the number of protein-protein interaction entries accounts of about 25% of the total number of entries in our Kinetic Data of Biomolecular Interaction (KDBI) database. As these entries provide kinetic data of protein-protein interactions, such a network is particularly useful for quantitative as well as qualitative study of biological pathways.

4.2.2 Procedure of protein-protein interaction network construction

The basic idea to construct the protein-protein interaction network is to connect a number of single "nodes" to form the network. Each "node" in the protein-protein interaction network is actually a unique protein molecule. The definition of unique protein in network construction is based on following rules: (1) proteins having same genetic identity such as name, function and sequence, but coming from different species are considered as one same protein, although these proteins give different kinetic values when they bind to the same molecule. (2) Wide-type protein and its mutants can be treated as one molecule, although they normally show different binding abilities binding to same molecule. For example, alcohol dehydrogenases from Homo sapiens and from rats will be served as same "node" in the network; wild type alcohol dehydrogenases and their mutants are also considered as same "node". A *PERL* program is written to re-identify the proteins in KDBI. At the same time, a protein-protein interaction table is generated automatically by the program to store the network information. An example of the table is shown in Table 4.8. Each record of this table consists of two parts of information: the left side of symbol "<-->" is the unique protein ID of the "starting node", while the right side part is the ID list of the proteins which interact with this specific protein, so-called the "ending nodes". It is noted that each protein on the right side of entries also acts as a "starting node" of another entry in the table. By linking the "starting node" to each protein of the "ending nodes", each entry of the interaction Table 4.8 forms a small protein-protein interaction "tree". Following the same method, a number of such small interaction "trees" are thus generated. Binding these small "tree" together through the common "nodes", the protein-

protein interaction network is constructed. Due to long and unformatted protein names, the network is constructed using unique molecule IDs instead of molecule names for better layout. However, to illustrate the relationship between the proteins, the mapping of molecule IDs to corresponding proteins is also listed following the protein network. The kinetic information of protein-protein interaction can be acquired by clicking the connecting lines between two proteins, which will pass the respective proteins as keywords to the KDBI search engine.

4.2.3 Result and analysis of the protein-protein interaction network construction

An example result of the network is shown at the Figure 4.6. The protein interaction network is shown at different levels. For example, the molecule, Leukocyte elastase PN00440 (start node, and level 1 of network), interacts with 5 molecules PN00439, PN00462, PN0463, PN00498 and PN00714 (level 2). And further, PN00463 interacts with 11 proteins: PN00115, PN00123, PN00128, PN00169, PN00225, PN00296, PN00465, PN00468, PN00480, PN00523, and PN00530 (level 3); Molecule PN00128 (Neutrophil elastase) further interacts with molecule PN00498 (level 2 or level 4); Molecule PN00480 further interacts with molecules PN00714 (level 2 or level 4) and PN00296 (level 4). For the 5 molecules at level 2 of network, three of them PN00463, PN00498 and PN00714 continue the network to level 3, the other two molecules terminate the interaction due to the limitation of the data. It is noted that there exist a loop interaction among molecules Leukocyte elastase (PN00440), Alpha-1-antitrypsin F (PN00463), Neutrophil elastase (PN00128) and Alpha-1-antitrypsin M (PN00498). Another loop is among proteins Alpha-

1-antitrypsin F (PN00463), Miniplasmin (PN00480) and Plasmin (PN00296). The loop interactions in the network are often related to non-leaner bio-events, which play important role in the study of protein profiling and regulation. The construction of the protein-protein interaction network is still preliminary. With more data available in KDBI, along with the improvement of visualization application in this project, the network will become more complete and user-friendly. It is expected combining with the existing metabolic pathway or regulatory pathway maps and that protein-protein interaction network may fill up the imaginary connection in the pathway maps, as well, the pathway maps will endow the interaction network with a broad view of biological processes.

Other than the protein-protein interaction network, protein-ligand interaction network based on KDBI is also a promising application for drug discovery. By studying the affinity of single protein-mulitple ligands interactions, one can develop a view of what ligands may interact with a protein. Such structure and activity relationship (SAR) study is recognized as a feasible approach for drug discovery. On the other hand, by studying the single ligand-multiple proteins interactions, one could explore the possibility of ligands as drug candidates: Does a ligand bind to expected therapeutic targets? Can it be transported to the therapeutic site by ADME associated proteins? Will it lead to serious side effects by interaction with targets?

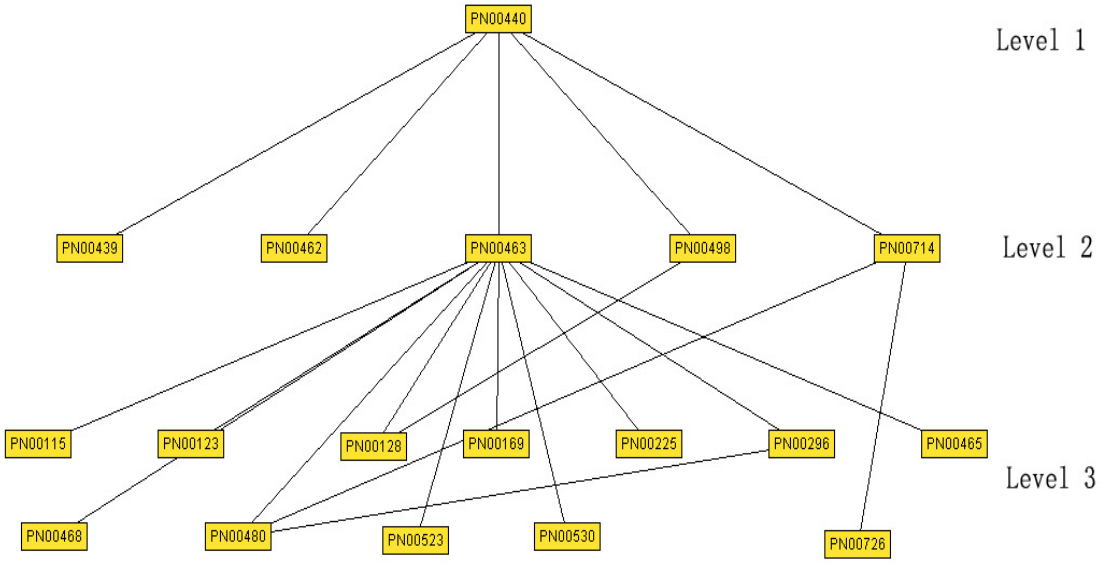
Another potential application of KDBI is the mathematical simulation of biological pathways. The availability of kinetic information enables the mathematical modeling of bio-molecular interactions. The modeling is done to further improve the understanding of cellular processes or biological pathways and identify potential therapeutic applications.

Similar achievement have been published [Grossniklaus *et al.*, 1996; Fussenegger *et al.*, 2000].

Table 4.8. The protein-protein relationship

Protein ID<-->Interacting protein IDs
PN00133<-->PN00892;PN00894
PN00142<-->PN01010;PN01012;PN01205;PN00861;PN00894
PN00146<-->PN00939;PN00894;PN00951;PN00952;PN00856;PN01320;PN00868
PN00149<-->PN01009;PN00183;PN00185;PN01052;PN01206
PN00150<-->PN01250;PN01252

Figure 4.6. The example of protein-protein interaction network.



PN00440-->Leukocyte elastase
PN00439-->Antichymotrypsin
PN00462-->Recombinant alpha1-abtitrypsin
PN00463-->Alpha-1-antitrypsin F
PN00498-->Alpha-1-antitrypsin M
PN00714-->Alpha 2-Macroglobulin
PN00115-->Protein C
PN00123-->Urokinase
PN00128-->Neutrophil elastase
PN00169-->Tissue plasminogen activator
PN00225-->Bovine trypsin
PN00296-->Plasmin
PN00465-->Kallikrein
PN00468-->Porcine pancreatic elastase
PN00480-->Miniplasmin
PN00523-->Sperm protease acrosin
PN00530-->Furin
PN00726-->Fibrin-bound miniplasmin

CHAPTER 5 CONCLUSION

5.1 Integration Of Subject-Specialized Databases for Comprehensive Information

In the early days, when biological databases were first opened for public access, the general opinion was doubtful and hesitant. Now, it has become a routine step for the researchers to make use of the biological databases to answer the questions before the expensive experiments are carried out. The biological databases ranges from the large-scale primary archiving projects such as GenBank and SWISS_PROT, to the individual and subject-specialized databases such as our DART and KDBI. Database systems today are facing the task of organizing the ever-increasing amount and complexity of biological data in the demand of different user communities.

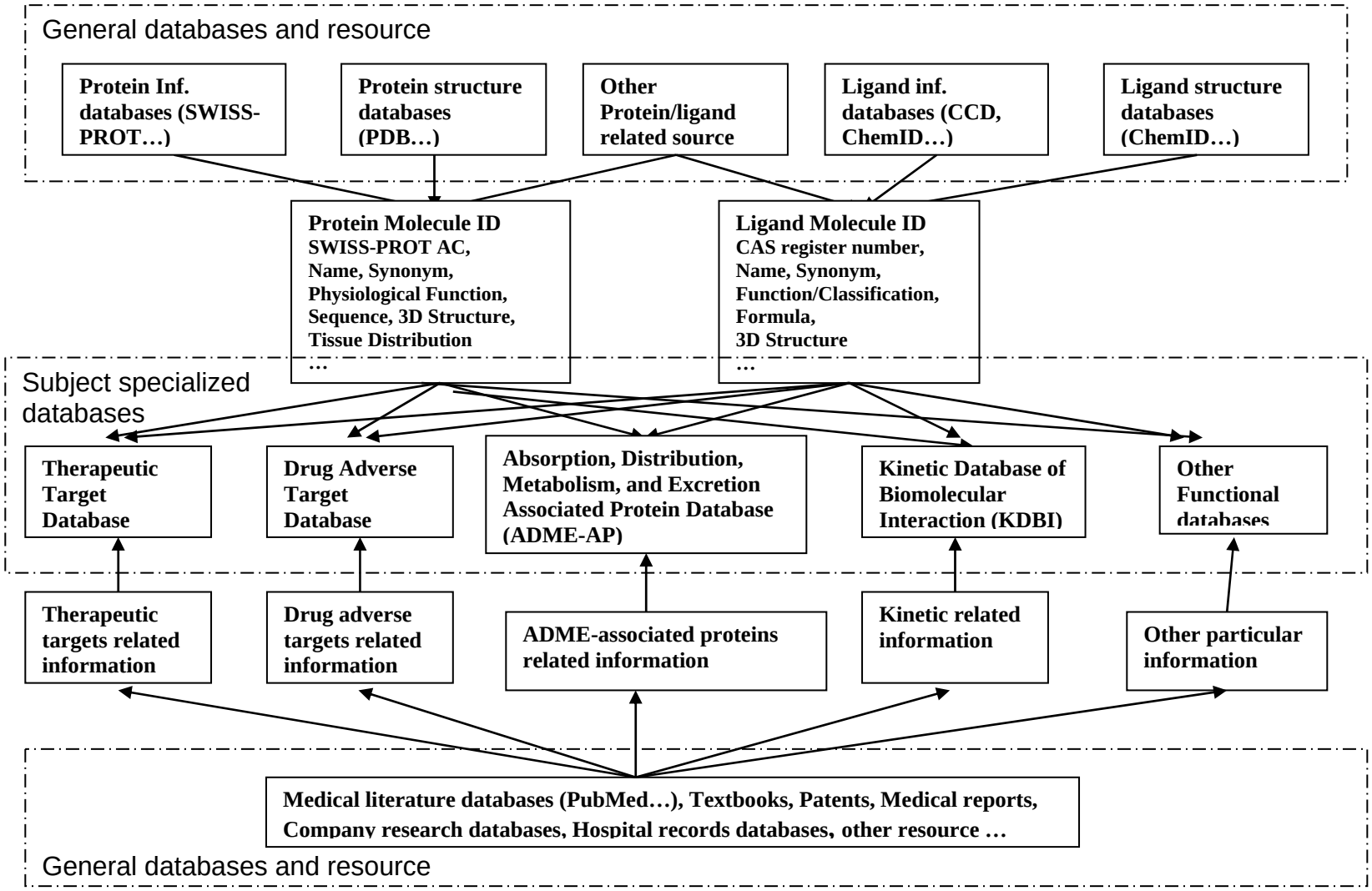
The general databases are normally large in size and contain raw data of sequence, structure, or literature. It is thought that they are databases rather than knowledge bases, which describe miscellaneous objects according to the database schema, but no representation of general concepts and their relationship [Hafner *et al.*, 1996]. As a result, further efforts are required to interpret the raw data so that they can be practically understood and used by communities having different interests. To meet the need of different communities, the subjected-specialized databases are constructed. Compared to the general databases, they are normally compact in size and focused on providing information about certain biological fields. The subject-specialized databases collect the subject related information from different data source on purpose. For example, the drug

adverse reaction target database (DART) only collects the information concerning the drug adverse effects and organizes them in context; such information can be found in different general sources that are isolated.

Users may be discouraged when they attempt to further explore some interested concepts in subject-specialized databases. A solution is to build the cross-links to other general or subjected-specialized databases containing concepts of interests. The process is called integration of databases. The integration of databases enables the subject-specialized databases to be compact, however, it provides more comprehensive information. For example, KDBI database cross-links to the general protein database SWISS_PROT for the additional protein properties such as sequence similarity; it also cross-links to the PDB for protein 3D structure information. The integration is achieved by sending keywords to the search engines or retrieval systems of respective databases. This level of integration is indirect and sometimes contains multiple intermediated steps. There is another level of data integration, which is direct and compact. The integration involves different local databases, which share groups of information. Normally, the integration process is taking place during the initial construction of databases. At that stage, database developers should determine which information components may be linked to other databases, and how table(s) storing common information components be shared in different databases. Any update of these common table(s) will affect all the connected databases so that it saves the workload for the overlap information. For example, as shown in Figure 5.1, several subject-specialized databases such as DART, KDBI, Therapeutic Target Database (TTD) [Chen *et al.*, 2002], and ADME associated protein database (ADME-AP) [Sun *et al.*, 2002] can share the same ligand information table since they require similar

information of ligands. And information of protein properties is collected from different source, but deposit in one table, which is also shared by different databases.

Figure 5.1. The construction of subject specialized databases



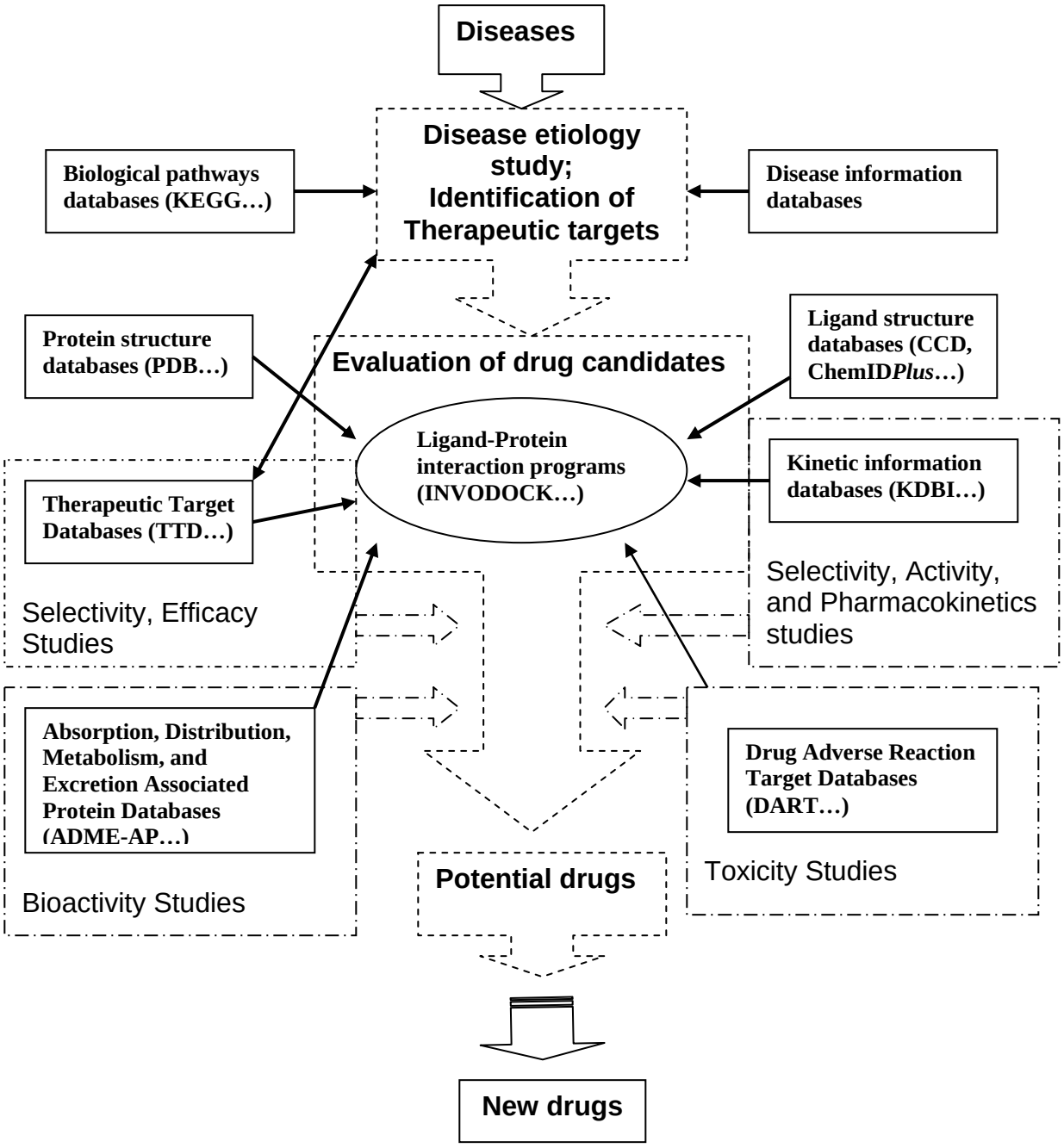
5.2 Proposal of a New CADD Approach: Drug Target Databases Aiding Drug Discovery

Protein targets play a crucial role in the disease etiology studies, pharmacokinetics studies or toxicity studies. Identification of these target proteins may help to enhance the efficacy of drugs and reduce the side effects. At least, three kinds of target proteins are important in drug discovery: therapeutic targets, ADME (absorption, distribution, metabolism, and excretion) associated proteins, and adverse drug reaction (ADR)/toxicity targets. In this work, the adverse drug reaction target database (DART) is created to help in the identification of potential toxicity targets and hence filter out the serious toxicity inducing drug candidates. The other two target databases therapeutic target databases TTD [Chen *et al.*, 2002] and ADME-associated protein databases ADME-AP [Sun *et al.*, 2003] were also previously developed by our research group.

These three target databases participate the computer-aided drug design at different stages. As shown in Figure 5.2, TTD help the selection of therapeutic targets during the disease etiology study and it is also involved in the efficacy study. The compounds in the pools are filtered by binding studies on the therapeutic targets. ADME-AP database will further help to remove those candidates with delivery or metabolism problems. Following efficacy, selectivity and bioactivity studies, the potential drugs candidates will be passed to toxicity study, which is supported by the DART databases. These databases can be combined with computer-aided molecular modeling software such as the docking program INVODOCK to facilitate drug discovery virtually. An example is the drug safety evaluation study described in Chapter 3, where combinational applications of

INVODOCK and DART database successfully predicted adverse effects induced by investigational drugs. Similar application can also be used to predict the efficacy, bioactivity, and pharmacokinetic characteristics of potential drugs by including inclusion of specific databases such as TTD, ADME-AP and KDBI. In conclusion, databases supported CADD system may be a good method to facilitate drug discovery. However, it is also noted that the robustness of this system greatly depends on the completion of databases. As the result, continuous update of these databases may be the key factor for practical application of this computer-aided drug design system.

Figure 5.2. Databases in drug discovery



5.3 Proper Prediction of ADR Targets by SVMs

Based on algorithm of Support Vector Machines (SVMs) and its applications in the proteomics and genomics studies, it appears that SVMs can be very useful in mining unknown patterns or features based on the primary sequence of proteins or nucleic acids. This view is further supported by our study on the prediction of ADR target proteins using SVMPROT. The overall accuracy of about 90% in the ADRs proteins prediction may prompt the identification of ADR targets. However, it is also noted that the accuracies for the prediction of positive samples and negative samples are different, 96.8% and 87.4% for positive and negative prediction, respectively.

The different performance of SVM on prediction and exact classification of positive and negative data is essentially due to the definition of positive and negative data for the learning model building. SVM learns to be optimal hyperplane classifier and able to separate the positives and negatives based on the computation of certain vectors (descriptors). If all the features describing the positive/negative data were selected to construct the hyperplane, the prediction accuracy of positives/negatives should reach close to 100% after optimization. However in practical, this may be difficult. One reason is the inability to determine all the essential descriptors for the proteins. An alternative solution is to include all possible descriptors for these proteins. The advantage is avoiding information loss; on the other hand, it also increases noise-to-signal ratios and the computational time. SVM optimization procedure and feature vector selection algorithm may also be improved by adding additional constraints and by incorporating independent component analysis and kernal functions such as PCA in the preprocessing steps.

Compared to the shortcoming of SVMs, the definition of data is also critical for the accuracy of classification. The definition of data is the problem of how to prepare the positive and negative datasets, especially for training and testing. Well-prepared positive/negative data will significantly increase the accuracy of SVM predictions and compensable for inadequate protein descriptors. A meaningful preparation should consist of two elements: firstly, all features for the separation of positives/negatives should be presented; and secondly, the pre-definition of positives/negatives should be correct and unambiguous. The first element covers the incomplete collection of the protein types. If one protein type with distinct features is not used for the model construction (here, the construction of hyperplane in SVM learning process), it will be very likely that all proteins of this type cannot be properly classified. This problem may be largely responsible that for the low accuracy of positives in our ADRs target proteins prediction. The incomplete collection of ADR target proteins significantly affects the performance of SVMs. However, it is believed that with more ADRs protein types found, the accuracy will increase. The second element is also important since incorrect assignment of positives and negatives may lead to incorrect hyperplanes for separation of input data. This will affect the accuracy of predicting both positives and negatives. A possible solution is to remove ambiguous data from the positive/negative datasets or reassign. A repetitive training process has positive effect on the accuracy of classification.

Other than the problems of descriptors and definition of the data, there are some more factors also affecting the performance of SVMs such as the size of data for training, the

protein sequence complement. Nevertheless, SVMs is a promising and plausible new approach for protein classification.

5.4 Information Extraction from Biomedical Literature Using Text Mining

Biomedical literature plays an important role in integrating, annotating and communicating experimental results and their implications. The information encapsulated in the literature is high and diverse, however, the distribution of the information is often diffuse and unrecognizable. In recent years, text mining technologies have been introduced into the life science to identify patterns of information and rules of information retrieval. In this work, two databases, DART and KDBI, are constructed by extracting the useful information from the public literature sources, and the text mining technology is also studied to help with information collection.

The semantic approach is common method in text mining, which can be realized at three levels: words, keywords and concepts. In this work, two lists of keywords are manually assigned for the feature selection: one is the keyword list of adverse effects and the other one is the distinct protein list. The ADR keywords are manually collected from the clinical reports. The list of protein keywords is generated by removing the redundant entries from SWISS_PROT protein list. Compared to the automatical identification of keywords, the manually assigned keywords are more accurate and practical. Therefore, the list of adverse reaction keywords is collected manually. Two rules are applied to constrain the abstracts: the abstract should include both the ADR keywords and protein, and the keywords should

be closely related. In this study, the relationship of the keywords was simply identified by the distance between the keywords. It is assumed that two keywords are related if they appear in the same sentence or an adjacent sentence. Thus a maximum distance is used to check the relationship between two keywords. This maximum distance is an empirical value obtained by many times of text mining. Unfortunately, keywords close in distance doesn't always mean the direct relationship between them. More restricted conditions may be helpful to filter out the unrelated abstracts containing both ADR and protein keywords. A better solution is to use concept representation instead of keyword representation to extract information. It is reasonable to believe that the evolution of text mining from words and keywords to concepts will greatly improve the extraction of truly useful information from literatures. However, further improvement of text mining algorithm is not carried out in this study due to consideration of research cost. The identification of the ADR target related information is difficult; therefore, any useful information-containing abstract is critical for the completeness of database. To protect from loss of useful information and provide correct data, manual editing of selected abstracts after simple keyword text mining is carried out.

The kinetic information of KDBI is collected completely manually. The reason for manual collection of information, instead of automatically mining by robot, is because of the insufficient electronic papers containing kinetic information. Unlike the information concerning adverse drug reactions, where the clues (keywords) and even the information itself often appear in the abstracts, generally, the majority of kinetic values are available only in the full papers in the forms of tables or lists. Without processing the full papers, much useful kinetic information will be missing, thus a nearly complete collection of

kinetic values becomes impossible. Unfortunately, the full papers in many journals are only available for a fee; moreover, the papers in electronic version feasible for text mining are limited in volume (normally, the volumes after year 1990s). Given that many kinetic values are listed in tables, reaction equations or figures, it would be harmful if the robots cannot properly determine which table contains kinetic information and what the useful kinetic information in the table is. Especially for those kinetic values appearing in the reaction equations, the robot extraction of information is almost impossible. The difficulties in obtaining the electronic version of full papers and efficiently extracting useful information from papers lead to impossible of using robots to gather kinetic information. The decision is also supported by the difficulty of verification of kinetic data generated by text mining. The verification of data is critical for a good publicly accessible database. The one-by-one checking process is always as time-consuming as manual information collection. Considering all conditions of information collection, manual collection is adopted finally.

In conclusion, text mining is promising in information extraction from the flood of documents. However, an absolutely depend on text mining to collect information from biological documents is impractical, especially in the development of public databases. The difficulties come both from the limitation of natural language recognition and the characteristics of biological fields, e.g., the very unstructured nature of biomedical documents, the confusion of biological nomenclature, and the diversity of biological information. The shortcomings of biomedical information sources are a historical problems, and little work can be done to change the situation. It is expected that some control measures in biological/medical fields such as the standardization of the protein

nomenclature can be accepted by all researchers in the future. Then information will be easier and more accurately accessible through text mining or other robot technologies.

REFERENCE

Ahonen H, Heinonen O, Klemettinen M, and Verkamo AI. Applying Data Mining Techniques in Text Analysis. Report C-1997-23, 1997, University of Helsinki, Department of Computer Science, Finland.

Ahonen H, Heinonen O, Klemettinen M, and Verkamo AI. Mining in the phrasal frontier. In Proceedings of PKDD'97-1st European Symposium on Principles of Data Mining and Knowledge Discovery, 1997, Norway.

Albers B, Bray D, Lewis J, Raff M, Roberts K and Watson JD. Molecular Biology of the Cell. 2nd edition. 1989. Garland, New York.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997 Sep 1; 25(17): 3389-3402.

Appelt D, Hobbs J, Bear J, Israel D, and Tyson M. FASTUS: A Finite-state Processor for Information Extraction from Real-world text. In Proc 13th Int'l Joint Conf Artificial Intelligence (IJCAI-93) 1993: 1172-1178.

Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 2003 Jan 1; 31(1): 248-250.

Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. BIND--The Biomolecular Interaction Network Database. Nucleic Acids Res 2001; 29(1): 242-245.

Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000 Jan 1; 28(1): 45-48.

Bairoch A, Bucher P. PROSITE: recent developments. Nucleic Acids Res 1994 Sep; 22(17): 3583-3589.

Bairoch A. The ENZYME database in 2000. Nucleic Acids Res. 2000 Jan 1; 28(1): 304-305.

Baldi P, Brunak S, Chauvin Y, Anderson CAF & Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000, 16, 412-419.

Barratt MD. Integrating Computer Prediction Systems With In Vitro Methods Towards A Better Understanding Of Toxicology. Toxicol Lett 1998; 102-103, 617-621.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002 Jan 1; 30(1): 276-280.

Baxevanis AD. The Molecular Biology Database Collection: 2003 update. Nucleic Acids Res 2003; 31(1): 1-12.

Baynes J, and Dominiczak MH. Medical Biochemistry, 1999, St Louis, Mosby.

Bellazzi R, Magni P, Larizza C, De Nicolao G, Riva A, Stefanelli M. Mining biomedical time series by combining structural analysis and temporal abstractions. Proc AMIA Symp 1998: 160-164.

Bellazzi R, Zupan B. Intelligent data analysis--special issue. Methods Inf Med 2001; 40(5): 362-4.

Benovic JL. Regulation of G Protein Coupled Receptor Function and Expression: Receptor Biochemistry and Methodology, 1999, Wiley-Liss.

Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res 1999 Jan 1; 27(1): 12-17.

Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR, and Schneider B. The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. Biophys J 1992; 63: 751-759.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000 Jan 1; 28(1): 235-242.

Bikel D, Miller S, Schwartz R and Weischedel R. Nymble: a high-performance learning name-finder. In Proc. Fifth Applied Natural Language Processing Conf, 1997, Washington, DC.

Black JW. Basic drug research in universities and industry. Br J Clin Pharmacol 1986; 22: 5-7.

Blaney JM and Hansch C. Use of molecular graphics for structural analysis of small molecules. Comp Med Chem 1990; 4: 459-496.

Bock JR, Gough DA. Predicting protein--protein interactions from primary structure. Bioinformatics 2001 May; 17(5): 455-460.

Boguski MS, McIntosh MW. Biomedical informatics for proteomics. Nature. 2003 Mar 13; 422(6928): 233-237.

Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997 Apr 25; 268(1): 78-94.

Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. 1998. Kluwer Academic Publishers, Boston, USA.

Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 2003 Jul 1; 31(13): 3692-3697.

Carazo JM, Stelzer EH, Engel A, Fita I, Henn C, Machtynger J, McNeil P, Shotton DM, Chagoyen M, de Alarcon PA, Fritsch R, Heymann JB, Kalko S, Pittet JJ, Rodriguez-Tome P, Boudier T. Organising multi-dimensional biological image information: the BioImage Database. Nucleic Acids Res. 1999 Jan 1; 27(1): 280-283.

Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. Nucleic Acids Res 2002 Jan 1; 30(1): 412-415.

Chen X, Liu M, Gilson MK. BindingDB: a web-accessible molecular recognition database. Comb Chem High Throughput Screen. 2001 Dec; 4(8): 719-725.

Chen YZ, and Ung, CY. Prediction Of Potential Toxicity And Side Effect Protein Targets Of A Small Molecule By A Ligand-Protein Inverse Docking Approach. J Mol Graph Mod 2001; 5278, 1-20.

Chen YZ, Zhi DG. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins 2001; 43(2): 217-226.

Chung SY, Wong L. Kleisli: a new tool for data integration in biology. Trends Biotechnol 1999 Sep; 17(9): 351-355.

Clerkin P, Cunningham P, and Hayes C. Ontology Discovery for the Semantic Web Using Hierarchical Clustering. Ontology Discovery for the Semantic Web. Workshop at ECML/PKDD, 2001.

Cottes C and Vapnik V. Support Vector Networks. Machine Learning 1995, 20: 273-297.

Coulter DM, Bate A, Meyboom RH, Lindquist M, Edwards IR. Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study. BMJ 2001 May 19; 322(7296): 1207-1209.

Cowie AP. English Dictionaries for the Foreign Learner. In Hartmann RRK ,editor, 1983,135-144.

Cunningham H, Gaizauskas RJ and Wilks Y. A General Architecture for Text Engineering (GATE)-A New Approach to Language Engineering R&D. Technical Report

CS-95-21, 1995, Institute for Language, Speech and Hearing (ILASH), Department of Computer Science, University of Sheffield, UK.

Cvetkovic RS, Goa KL. Lopinavir/ritonavir: a review of its use in the management of HIV infection. Drugs. 2003; 63(8): 769-802.

Dayhoff MO, Schwartz RM, Chen HR, Hunt LT, Barker WC, Orcutt BC. Nucleic acid sequence bank. Science 1980; 209(4462): 1182.

DeJong G. An Overview of the FRUMP system. Strategies for Natural Language Processing. In Lehnert WB and Ringle MH, editors, 1982, 149-176.

Denny BJ, Wheelhouse RT, Stevens MF, Tsang LL, Slack JA. NMR and molecular modeling investigation of the mechanism of activation of the antitumor drug temozolomide and its interaction with DNA. Biochemistry 1994 Aug 9; 33(31): 9045-9051.

Dixon M. An Overview of Document Mining Technology. 1997, the Australian National University, Australia.

Downward J. The ins and outs of signaling. Nature 2001; 411(6839): 759-762.

Dusseldorp E, Meulman JJ. Prediction in medicine by integrating regression trees into regression analysis with optimal scaling. Methods Inf Med 2001; 40(5): 403-409.

Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. Methods Enzymol 1996; 266: 114-128.

Feldman R and Dagan I. Knowledge discovery in textual databases (kdt). In Proceedings of the First International Conference on Knowledge Discovery (KDD-95). ACM, August 1995.

Fradera X, Knegtel RM, Mestres J. Similarity-driven flexible ligand docking. Proteins 2000; 40(4): 623-636.

Frawley W, Piatetsky-Shapiro G, and Matheus C. Knowledge Discovery in Databases: An Overview. AI Magazine 1992, 13(3): 57-70.

Frishman D, Heumann K, Lesk A, Mewes HW. Comprehensive, comprehensible, distributed and intelligent databases: current status. Bioinformatics 1998; 14(7): 551-561.

Fritz TA, Tondi D, Finer-Moore JS, Costi MP, Stroud RM. Predicting and harnessing protein flexibility in the design of species-specific inhibitors of thymidylate synthase. Chem Biol 2001; 8(10): 981-995.

Fullton J. WAIS. Interlligent Information Retrieval: The case of Astronomy and Related Space Sciences 1993, in Heck A and Murtagh F editor, Kluwer, Dordrecht, 113-118.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000 Oct; 16(10): 906-914.

Fussenegger M, Bailey JE, Varner J. A mathematical model of caspase function in apoptosis. Nat Biotechnol. 2000 Jul; 18(7): 768-774.

George DG, Mewes HW, Kihara H. A standardized format for sequence data exchange. Protein Seq Data Anal 1987; 1(1): 27-39.

Gerhold D, Rushmore T, And Caskey CT. DNA Chips: Promising Toys Have Become Powerful Tools. Trends Biochem Sci 1999; 24: 168-173.

Gleich LL, Collins CM, Gartside PS, Gluckman JL, Barrett WL, Wilson KM, Biddinger PW, Redmond KP. Therapeutic decision making in stages III and IV head and neck squamous cell carcinoma. Arch Otolaryngol Head Neck Surg 2003 Jan; 129(1): 26-35.

Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: applications of AutoDock. J Mol Recognit 1996; 9(1): 1-5.

Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K, Kanehisa M. Organizing and computing metabolic pathway data in terms of binary relations. Pac Symp Biocomput 1997: 175-186.

Goto S, Nishioka T, Kanehisa M. LIGAND: chemical database for enzyme reactions. Bioinformatics 1998; 14(7): 591-599.

Greller LD, Tobin FL. Detecting selective expression of genes and proteins. Genome Res 1999 Mar; 9(3): 282-296.

Grishaman R. New York University PROTEUS System: MUC-4 Test Results and Analysis. In Proceedings of the Sixth Message Understanding Conference (MUC-6): 124-127, 1995, San Mateo, Morgan Kaufmann.

Grishman R. Information extraction: Techniques and challenges. Information Extraction: A multidisciplinary Approach to an Emergine Information Technology, 1997, 1299: 10-27.

Grossniklaus U, Madhusudhan MS, Nanjundiah V. Nonlinear enzyme kinetics can lead to high metabolic flux control coefficients: implications for the evolution of dominance. J Theor Biol. 1996 Oct 7; 182(3): 299-302.

Hafner CD, Fridman N. Ontological foundations for biology knowledge models. Proc Int Conf Intell Syst Mol Biol 1996; 4: 78-87.

Haigh T. A Veritable Bucket of Facts: Origins of the Database Management System. In Proceedings of the Second Conference on the History and Heritage of Scientific and Technical Information Systems, Information Today, Medford, 2003.

Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002; 47(4): 409-443.

Han JW, Fu YJ. Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. KDD Workshop 1994: 157-168.

Haney SA, Alksne LE, Dunman PM, Murphy E, Projan SJ. Genomics in anti-infective drug discovery--getting to endgame. Curr Pharm Des 2002; 8(13): 1099-1118.

Hansch C, Hoekman D, Leo A, Zhang L, Li P. The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. Toxicol Lett 1995 Sep; 79(1-3): 45-53.

Hart GJ, Orr DC, Penn CR, Figueiredo HT, Gray NM, Boehme RE, Cameron JM. Effects of (-)-2'-deoxy-3'-thiacytidine (3TC) 5'-triphosphate on human immunodeficiency virus reverse transcriptase and mammalian DNA polymerases alpha, beta, and gamma. Antimicrob Agents Chemother. 1992 Aug; 36(8): 1688-1694.

Haugh JM, Wells A, Lauffenburger DA. Mathematical modeling of epidermal growth factor receptor signaling through the phospholipase C pathway: mechanistic insights and predictions for molecular interventions. Biotechnol Bioeng 2000; 70(2): 225-238.

Hearst MA, Scholkopf B, Dumais S, Osuna E, and Platt J. Trends and controversies-support vector machines. IEEE Intelligent Systems 1998, 13, 18-28.

Hopfinger AJ. Computer-assisted drug design. J Med Chem 1985 Sep; 28(9): 1133-1139.

Horn F, Weare J, Beukers MW, Horsch S, Bairoch A, Chen W, Edvardsen O, Campagne F, Vriend G. GPCRDB: an information system for G protein-coupled receptors. Nucleic Acids Res 1998 Jan 1; 26(1): 275-279.

Huang X, Moy F, Powers R. Evaluation of the utility of NMR structures determined from minimal NOE-based restraints for structure-based drug design, using MMP-1 as an example. Biochemistry 2000; 39(44): 13365-13375.

Igarashi T, Kaminuma T. Development of a cell signaling networks database. Pac Symp Biocomput 1997: 187-197.

Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 1999: 138-148.

Jacques FM, Linden L and Selby N. Turning the pipeline into a hit parade. Pharm Exec 1992; 12 (3): 62-72.

Jakobsen IB, Saleeba JA, Poidinger M, Littlejohn TG. TreeGeneBrowser: phylogenetic data mining of gene sequences from public databases. Bioinformatics 2001 Jun; 17(6): 535-540.

Ji ZL, Chen X, Zhen CJ, Yao LX, Han LY, Yeo WK, Chung PC, Puy HS, Tay YT, Muhammad A, Chen YZ. KDBI: Kinetic Data of Bio-molecular Interactions database. Nucleic Acids Res. 2003 Jan 1; 31(1): 255-257.

Ji ZL, Han LY, Yap CW, Sun LZ, Chen X, Chen YZ. Drug Adverse Reaction Target Database (DART): Proteins Related to Adverse Drug Reactions. Drug Saf. 2003; 26(10): 685-690.

Joly V, Descamps D, Yeni P. NNRTI plus PI combinations in the perspective of nucleoside-sparing or nucleoside-failing antiretroviral regimens. AIDS Rev. 2002 Jul-Sep; 4(3): 128-139.

Jonassen I, Eidhammer I, Conklin D, Taylor WR. Structure motif discovery and mining the PDB. Bioinformatics 2002 Feb; 18(2): 362-367.

Kakuda TN. Pharmacology of nucleoside and nucleotide reverse transcriptase inhibitor-induced mitochondrial toxicity. Clin Ther. 2000 Jun; 22(6): 685-708.

Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002 Jan 1; 30(1): 42-46.

Keiser J, Stich A, Burri C. New drugs for the treatment of human African trypanosomiasis: research and development. Trends Parasitol 2001 Jan; 17(1): 42-49.

Klaassen CD. Casarett & Doull's Toxicology: The Basic Science Of Poisons (6th Edition). New York: Mcgraw-Hill, 2001.

Knoblock CA, Lerman K, Minton S and Muslea I. A Machine Learning Approach to Accurately and Reliably Extracting Data from the Web. IEEE Data Engineering Bulletin 2000, 23(4): 33-41.

Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res. 2002 Jan 1; 30(1): 312-317.

Kong AN, Mandlekar S, Yu R, Lei W, And Fasanmande A. Pharmacodynamics And Toxicodynamics Of Drug Action: Signalling In Cell Survival And Cell Death. Pharmaceut Res 1999; 16: 790-798.

Koonin EV, Tatusov RL and Galperin MY. Beyond complete genomes: from sequence to structure and function. Curr Opin Struct Biol. 1998 Jun; 8: 355-363.

Krumrine J, Raubacher F, Brooijmans N, Kuntz I. Principles and methods of docking and ligand design. Methods Biochem Anal 2003; 44: 443-476.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982; 161(2): 269-288.

Lavrac N. Selected Techniques for Data Mining in medicine. Artif Intell Med 1999 May; 16(1): 3-23.

Lengeler JW. Metabolic networks: a signal-oriented approach to cellular models. Biol Chem 2000; 381(9-10): 911-920.

Liebman MN. Biomedical informatics: the future for drug development. Drug Discov Today. 2002 Oct 15; 7(20 Suppl): 197-203.

Lynn DJ, Lloyd AT, O'Farrelly C. Bioinformatics: implications for medical research and clinical practice. Clin Invest Med. 2003 Apr; 26(2): 70-74.

Marshall GR and Naylor CB. Use of molecular graphics for structural analysis of small molecules. Comp Med Chem 1990; 4: 431-458.

McEntyre J, Lipman D. Pubmed: Bridging The Information Gap. CMAJ 2001; 164(9): 1317-1319.

McMartin C, Bohacek RS. QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 1997; 11(4): 333-344.

Monks TJ, And Lau SS. The Pharmacology And Toxicology Of Polyphenolic-Glutathione Conjugates. Annu Rev Pharmacol Toxicol 1998; 38: 229-255.

Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995 Apr 7; 247(4): 536-540.

Nakata K, Takai T, Kaminuma T. Development of the receptor database (RDB): application to the endocrine disruptor problem. Bioinformatics 1999 Jul-Aug; 15(7-8): 544-552.

Nandi T, B-Rao C, Ramachandran S. Comparative genomics using data mining tools. J Biosci 2002; 27(1 Suppl 1): 15-25.

Narasimhan G, Bu C, Gao Y, Wang X, Xu N, Mathee K. Mining protein sequences for motifs. J Comput Biol 2002; 9(5): 707-720.

Neri F and Saitta L. Machine Learning for Information Extraction. Information Extration: A Multidisciplinary Approach to an Emergine Information Technology, 1997, 1299: 171-191.

Ng SK, Wong M. Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. Genome Inform Ser Workshop Genome Inform 1999; 10: 104-112.

Norel R, Petrey D, Wolfson HJ, Nussinov R. Examination of shape complementarity in docking of unbound proteins. Proteins 1999; 36(3): 307-317.

Nuwaysir EF, Bittner M, Trent J, And Barrrett JC. Microarrays And Toxicology: The Advent Of Toxicogenomics. Mol Carcinog 1999; 24: 153-159.

Oprea TI, Waller CL, Marshall GR. Three-dimensional quantitative structure-activity relationship of human immunodeficiency virus (I) protease inhibitors. 2. Predictive power using limited exploration of alternate binding modes. J Med Chem 1994; 37(14): 2206-2215.

Park BK, Kitteringham NR, Powell H, And Primohamed M. Advances In Molecular Toxicology – Towards Understanding Idiosyncratic Drug Toxicity. Toxicology 2000; 153: 39-60.

Park BK, Pirmohamed M, Tingle MD, Madden S and Kitteringham NR. Bioactivation and bioinactivation of drugs and drug metabolites: relevance to adverse drug reactions. Toxicol In Vitro 1994; 8: 613-621.

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988 Apr; 85(8): 2444-2448.

Peltonen L and Mckusick VA. Genomics And Medicine: Dissecting Human Disease In The Postgenomics Era. Science 2001; 291: 1224-1232.

Persidis A. Bioinformatics. Nat Biotechnol 1999 Aug; 17(8): 828-830.

Pirmohamed M, Park BK. Genetic Susceptibility To Adverse Drug Reactions. Trends Pharmacol Sci 2001; 22(6): 298-305.

Politowska E, Drabik P, Kazmierkiewicz R, Ciarkowsk J. Docking ligands to vasopressin and oxytocin receptors via genetic algorithm. J Recept Signal Transduct Res 2002; 22(1-4): 393-409.

Pumford NR, and Halmes NC. Protein Targets Of Xenobiotic Reactive Intermediates. Annu Rev Pharmacol Toxicol 1997; 37: 91-117.

Ramakrishnan R and Gehrke J. Database Management Systems. 3nd edition. University of Winsconsin-Madison, 2002.

Rang HP, Dale MM, and Ritter JM. Pharmacology (4th Edition). New York: Churchill Livingstone, 1999.

Rosenquist M, Alsterfjord M, Larsson C, Sommarin M. Data mining the Arabidopsis genome reveals fifteen 14-3-3 genes. Expression is demonstrated for two out of five novel genes. Plant Physiol 2001 Sep; 127(1): 142-149.

Sager N, Friedman C, and Lyman M. Medical Language Processing: Computer Management of Narrative Data. Addision Wesley, 1987.

Sahm H, Eggeling L, de Graaf AA. Pathway analysis and metabolic engineering in Corynebacterium glutamicum. Biol Chem 2000; 381(9-10): 899-910.

Sali, A. 100,000 Protein Structures For Biologist. Nature Struct Biol 1998; 5: 1029-1032.

Sandak B, Wolfson HJ, Nussinov R. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. Proteins 1998; 32(2): 159-174.

Sarachan BD, Simmons MK, Subramanian P, Temkin JM. Combining Medical Informatics and Bioinformatics toward Tools for Personalized Medicine. Methods Inf Med. 2003; 42(2): 111-115.

Saunders J, Freedman SB. The design of full agonists for the cortical muscarinic receptor. Trends Pharmacol Sci 1989 Dec; Suppl: 70-75.

Schölkopf B, Burges C and Vapnik V. Extracting Support Data for a Given Task. In Fayyad UM and Uthurusamy R, editors, Proceedings, First International Conference on Knowledge Discovery & Data Mining 1995. AAAI Press, Menlo Park,CA.

Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res 2002 Jan 1; 30(1): 47-49.

Sese J, Nikaidou H, Kawamoto S, Minesaki Y, Morishita S, Okubo K. BodyMap incorporated PCR-based expression profiling data and a gene ranking system. Nucleic Acids Res. 2001 Jan 1; 29(1): 156-158.

Shi LM, Myers TG, Fan Y, O'Connor PM, Paull KD, Friend SH, Weinstein JN. Mining the National Cancer Institute Anticancer Drug Discovery Database: cluster analysis of

ellipticine analogs with p53-inverse and central nervous system-selective patterns of activity. Mol Pharmacol 1998 Feb; 53(2): 241-251.

Shoichet BK, Bodian DL, Kuntz ID. Molecular docking using shape descriptors. J Comp Chem 1992; 13: 380-397.

Sim E, Dimoglo A, Shvets N, Ahsen V. Electronic-topological study of the structure-activity relationships in a series of piperidine morphinomimetics. Curr Med Chem 2002; 9(16): 1537-1545.

Sivakumaran S, Hariharaputran S, Mishra J, Bhalla US. The Database of Quantitative Cellular Signaling: management and analysis of chemical kinetic models of signaling networks. Bioinformatics. 2003 Feb 12; 19(3): 408-415.

Smith LL. Key Challenges For Toxicologists In The 21st Century. Trends Pharmacol Sci 2001; 22(6): 281-285.

Sneader W. Chronology of drug introductions. Comp Med Chem 1990; 1: 7-80.

Stammberger I, Schmahl W, Tempel K. Scheduled and unscheduled DNA synthesis in chick embryo liver following X-irradiation and treatment with DNA repair inhibitors in vivo. Int J Radiat Biol. 1989 Sep; 56(3): 325-333.

Stoesser G, Moseley MA, Sleep J, McGowran M, Garcia-Pastor M, Sterk P. The EMBL nucleotide sequence database. Nucleic Acids Res 1998 Jan 1; 26(1): 8-15.

Sun LZ, Ji ZL, Chen X, Wang JF, Chen YZ. ADME-AP: a database of ADME associated proteins. Bioinformatics 2002; 18(12): 1699-1700.

Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr 1998 Nov 1; 54(Pt 6 Pt 1): 1078-1084.

Tantillo C, Ding J, Jacobo-Molina A, Nanni RG, Boyer PL, Hughes SH, Pauwels R, Andries K, Janssen PA, Arnold E. Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. Implications for mechanisms of drug inhibition and resistance. J Mol Biol 1994 Oct 28; 243(3): 369-387.

Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 2002 Jan 1; 30(1): 27-30.

Teeter MM, Froimowitz M, Stec B, DuRand CJ. Homology modeling of the dopamine D2 receptor and its testing by docking of agonists and tricyclic antagonists. J Med Chem 1994; 37(18): 2874-2888.

van Helden J, Naim A, Mancuso R, Eldridge M, Wernisch L, Gilbert D, Wodak SJ. Representing and analysing molecular and cellular function using the computer. Biol Chem 2000; 381(9-10): 921-935.

Vapnik V and Chervonenkis A. Theory of Pattern Recognition. 1974. Nauka, Moscow.
Vapnik V and Chervonenkis. A note on One Class of Perceptrons. Automation and Remote Control, 1964: 25.

Vapnik V and Lerner A. Pattern recognition using generalized portrait method. Automation and Remote Control, 1963: 24.

Vapnik V. Estimation of Dependences Based on Empirical Data. 1979. Nauka, Moscow, Russia.

Vapnik V. The Nature of Statistical Learning Theory 1995. Springer Verlag, New York
Vesell ES. Advances In Pharmacogenetics And Pharmacogenomics. J Clin Pharmacol 2000, 40: 930-938.

Wallace KB, And Starkov AA. Mitochondrial Targets Of Drug Toxicity. Annu Rev Pharmacol Toxicol 2000; 40: 353-388.

Wang J, Kollman PA, Kuntz ID. Flexible ligand docking: a multistep strategy approach. Proteins 1999; 36(1): 1-19.

Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002 Dec 5; 420(6915): 520-562.

Wender PA, Hinkle KW, Koehler MF, Lippa B. The rational design of potential chemotherapeutic agents: synthesis of bryostatin analogues. Med Res Rev 1999; 19(5): 388-407.

Wheatley M. Understanding neurotransmitter receptors: molecular biology-based strategies. Essays Biochem 1998; 33: 15-27.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 2003 Jan 1; 31(1): 28-33.

Williams M. Receptor binding in the drug discovery process. Med Res Rev 1991; 11(2): 147-184.

Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC, Suzek BE, Tsugita A, Vinayaka CR, Yeh LS, Zhang J, Barker WC. The Protein Information Resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Res. 2002 Jan 1; 30(1): 35-37.

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucleic Acids Res. 2000 Jan 1; 28(1): 289-291.

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002; 30(1): 303-305.

Zien A, Rätsch G, Mika S, Schölkopf B, Lemmen C, Smola A, Lengauer T and Müller KR. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. (1999). GCB '99, Hannover, Germany.

APPENDIX A

ALGORITHM OF SUPPORT VECTOR MACHINES

Support Vector Machine (SVM) is a learning technology introduced in 1979 [Vapnik *et al.*, 1979]. However, it receives increasing attention since it has been re-introduced by Dr. Burges [Burges *et al.*, 1998]. The principle of the SVM method is to learn from the related examples as a basis for predictions. In other words, SV algorithm is the connection between learning theory and practical applications. Generally, the algorithm of Support Vector Machines is composed of four stages: learning pattern recognition, hyperplane classification, kernel functions for feature spaces, and SV function estimation.

Learning Pattern Recognition

The start of support vector learning process is the problem of learning how to recognize patterns. A function f: $R^{\mathbf{N}} \rightarrow \{\pm 1\}$ is estimated using training data set $(\boldsymbol{x}_i, y_i)$ for pattern recognition. $\boldsymbol{x}_i$ are the N-dimensional patterns and y_i are the class labels, which are under the same probability distribution $P(\boldsymbol{x}, y)$,

$$(x_1, y_2),(x_2, y_2),...(x_l, y_l) \in R^N X\{\pm 1\} \qquad (1)$$

The function f is well generalized so that the training dataset $(\boldsymbol{x}_i, y_i)$, i = 1, 2, …, l, satisfy $f(\boldsymbol{x}_i) = y_i$. Through the learning, the function f should be able to correctly classify new examples $(\underline{\boldsymbol{x}}_i, \underline{y}_i)$, satisfying $f(\underline{\boldsymbol{x}}_i) = \underline{y}_i$. However, the fact is the well generalized function f from the training dataset doesn't have to be well generalized for the unseen new data. That is, for any test dataset $(\underline{\boldsymbol{x}}_i, \underline{y}_i) \in R^{\mathbf{N}}$ X $\{\pm 1\}$ and $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_j\} \cap \{\underline{\boldsymbol{x}}_1, \underline{\boldsymbol{x}}_2, \ldots, \underline{\boldsymbol{x}}_i\} = \{\ \}$, there exists another function f^* such that $f^*(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$ for all i = 1, 2, …, l, yet $f^*(\boldsymbol{x}_j) \neq f$

($\boldsymbol{x}_j$) for all j = 1, 2, …, l. Hence, there is no means to decide that which of these two functions is preferable, and only minimizing the training error thus does not imply a small test error. To minimizing the test error, it is good to restrict the class of functions that the machine learning can implement to one with a capacity that is suitable for the amount of available training data. The capacity is the ability of the SV machine to learn any training set without error. The statistical learning theory [Vapnik *et al.*, 1974; Vapnik *et al.*, 1979] or the VC (Vapnik-Chervonenkis) theory is thus introduced to add the bounds on the test error. The minimization of these bounds, which depend on both the empirical risk (training error) and the capacity of the function class, leads to the principle of structural risk minimization [Vapnik *et al.*, 1979]. The best-known capacity concept of VC theory is the VC dimension, defined as the largest number h of points that can be separated in all possible ways using functions of given class. If the $h < l$ is the VC dimension of the class of functions that the machine learning can implement, then for all functions of that class, the bound with a probability of at least 1- η will be

$$R(\alpha) \le R_{emp}(\alpha) + \phi(\frac{h}{l}, \frac{\log(\eta)}{l}) \tag{2}$$

where the confidence term ϕ is defined as

$$\phi(\frac{h}{l}, \frac{\log(\eta)}{l}) = \sqrt{\frac{h(\log\frac{2l}{h}+1) - \log(\frac{\eta}{4})}{l}}\,. \tag{3}$$

By the bound, for a finite amount of training dataset satisfying the distribution

$$P(x, y) = P(x) \cdot P(y)\,, \tag{4}$$

which means the pattern $\boldsymbol{x}$ contains no information about the label y, zero training error is possible. In order to reproduce the random labelings (increasing the capacity), a large VC

dimension *h* is required; the increase of *h* is accompanied by the increase of the confident term ϕ (4), thus the small test error will not be supported by the bound (3) and the accuracy is lowered. However, too little capacity will make the learning meaningless. Therefore, in order to get nontrivial predictions from (3), the function space must be restricted such that the capacity is small enough. For the given finite amount of training data set, there should exist a function having a balance between the classification accuracy and the capacity.

Hyperplane Classifiers

To design learning algorithms as well as finding the balance for machine learning between accuracy and the capacity, the capacity for a class of functions should be computed. Vapnik and Lerner [Vapnik *et al.*, 1963], and Vapnik and Chervonenkis [Vapnik *et al.*, 1964] proposed a learning algorithm for constructing the decision function *f* from empirical data for separating the class of hyperplanes. The class of hyperplanes is the base of SV classifier,

$$(w \cdot x) + b = 0 \qquad \boldsymbol{w} \in R^{\mathrm{N}}, \mathrm{b} \in R, \tag{5}$$

where $\boldsymbol{w}$ is the weight vector and the corresponding decision functions

$$f(x) = sign((w \cdot x) + b) \tag{6}$$

Among all hyperplanes separating the data, there exists a unique one, optimal hyperplane, yielding the maximum margin of separation between the classes,

$$\frac{\max}{w,b} \quad \min\{\| x - x_i \| : x \in \mathrm{R}^{\mathrm{N}}, (w \cdot x) + b = 0, i = 1,2,...,l\}, \tag{7}$$

and the capacity of the learning algorithm decreases while the margin increases (Figure 1) [Hearst *et al.*, 1998]. The construction of the Optimal Hyperplane is by solving the following optimization problem:

minimize $\tau(w) = \frac{1}{2} \| w \|^2$ (8)

subject to $y_i \cdot ((w \cdot x_i) + b) \geq 1, \quad i = 1, 2, \ldots, l.$ (9)

To solve the constrained optimization problem, the Lagrangian and the Lagrange multiplier α_i is introduced,

$$L(w, b, \alpha) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{l} \alpha_i (y_i \cdot ((x_i \cdot w) + b) - 1) \quad (10)$$

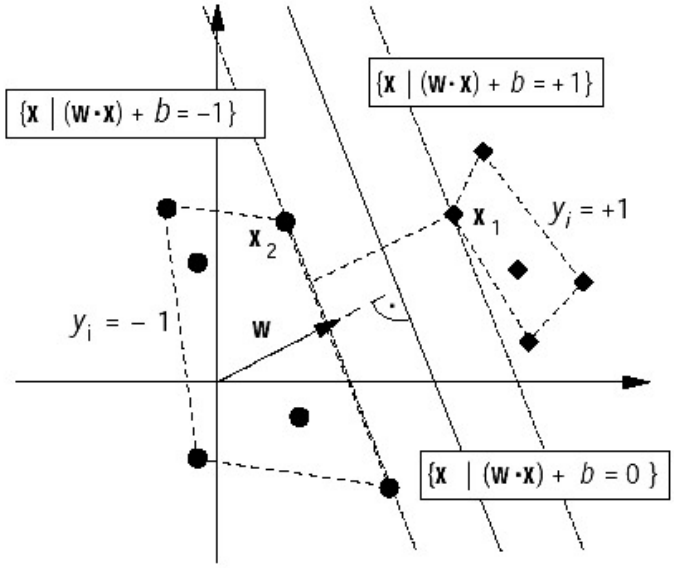
where $\alpha_i \geq 0$. The Lagrangian L has to be minimized with respect to the *primal variables* $\boldsymbol{w}$ and b and maximized with respect to the dual variables α_i. $\boldsymbol{w}$ here has an expansion $w = \sum_i \alpha_i y_i x_i$ in terms of a subset of the training patters, called *Support Vector* while α_i is non-zero. Solving the formula (10) subject to $\sum_{i=1}^{l} \alpha_i y_i = 0$ and $\alpha_i \geq 0$, the hyperplane decision function can thus be written as

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i \cdot (x \cdot x_i) + b) \quad (11)$$

where b is calculated by

$$\alpha_i \cdot [y_i ((x_i \cdot w) + b) - 1] = 0, \; i = 1, 2, \ldots, l. \quad (12)$$

Figure 1. The binary classification and the hyperplane.



Note:

$$(\mathbf{w} \cdot \mathbf{x}_1) + b = +1$$
$$(\mathbf{w} \cdot \mathbf{x}_2) + b = 1$$
$$\Rightarrow (\mathbf{w} \cdot (\mathbf{x}_1 - \mathbf{x}_2)) = 2$$
$$\Rightarrow \left(\frac{\mathbf{w}}{||\mathbf{w}||} \cdot (\mathbf{x}_1 - \mathbf{x}_2) \right) = \frac{2}{||\mathbf{w}||}$$

Feature Spaces and Kernels

To construct SV machines, the optimal hyperplane algorithm should allow a method for computing dot products in feature spaces nonlinearly related to input space. The basic idea is to map the data into some other dot product space, so-called the *feature space*, *F* via a nonlinear map,

$$\phi : R^N \rightarrow F \tag{12}$$

and perform the above linear algorithm in *F*. This requires the evaluation of dot products by a simple kernel function,

$$k(x, y) := (\phi(x) \cdot \phi(y)). \tag{13}$$

If *F* is high-dimensional, then kernel function, polynomial kernel,

$$k(x, y) = (x \cdot y)^d. \tag{14}$$

can be shown to correspond to a map ϕ into the space spanned by all products of exactly d dimensions of R^N. For example, d = 2 and $\boldsymbol{x}, y \in R^2$, then

$$(x \cdot y) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (\phi(x) \cdot \phi(y)), \tag{15}$$

defining $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$. For every kernel that gives rise to a positive matrix $(k(x_i, x_j))_{ij}$, a map ϕ can be constructed.

SV function estimation

To construct SV machines, one computes an optimal hyperplane in feature space. In practice, such a separating hyperplane may not exist due to high noise level of data. Thus the slack variables ξ are introduced. The modified classifier will generalized well by minimizing the objective function

$$\tau(w,\xi) = \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{l}\xi_i \tag{16}$$

with the relaxed constraints,

$$y_i \cdot ((w \cdot x_i) + b) \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, l. \tag{17}$$

and

$$\xi_i \geq 0, \qquad i = 1, 2, \ldots, l. \tag{18}$$

C is the upper bound on the Lagrange multipliers α_i,

$$0 \leq \alpha_i \leq C, \; i = 1, 2, \ldots, l, \text{ and } \sum_{i=1}^{l}\alpha_i y_i = 0. \tag{19}$$

The concept of the margin is specific to pattern recognition. To generalize the SV algorithm to regression estimation [Vapnik *et al.*, 1995], an analogue of the margin is constructed in the space of the target values y by using Vapnik's ε-insensitive loss function,

$$| y - f(x) |_{\varepsilon} := \max\{0, | y - f(x) | - \varepsilon\}. \tag{20}$$

In input space, the hyperplane corresponds to a nonlinear decision function, which is determined by the kernel (Figure 2) [Cortes *et al.*, 1995; Vapnik *et al.*, 1995]. By the choice of different kernel functions, different architectures can be achieved. Surprisingly, the different kernel functions lead to very similar classification accuracies and SV sets [Schölkopf *et al.*, 1995]. In this case, the learning algorithm tries to construct a linear function in the feature space such that the training points lie within a distance $\varepsilon > 0$. The algorithm can be modified such that ε need not be specified a priori. Instead, and an upper bound $0 \leq v \leq 1$ is specified on the fraction of points and the corresponding ε can be

computed automatically, where $v_i = yi\alpha_i$. Thus the optimization problem become solving the function,

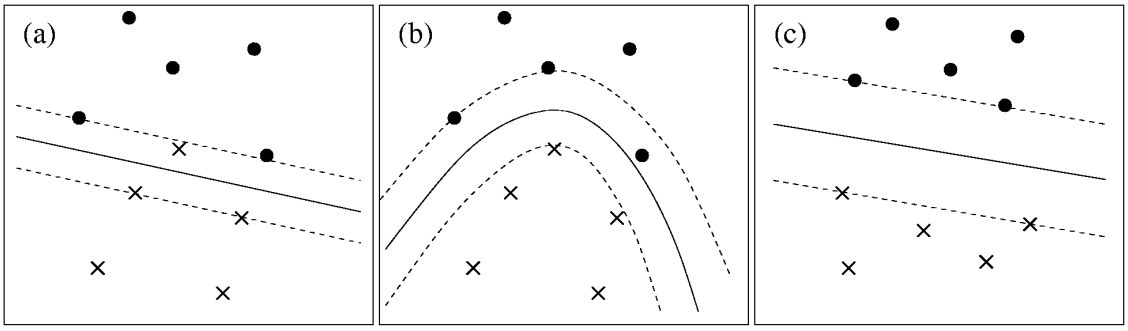
$$\frac{1}{2}\| w \|^2 + C\left(vl\varepsilon + \sum_{i=1}^{l} | y_i - f(x_i) |_{\varepsilon} \right) \quad (21)$$

Generally, the architecture of the learning process is illustrated in Figure 3.

Notation

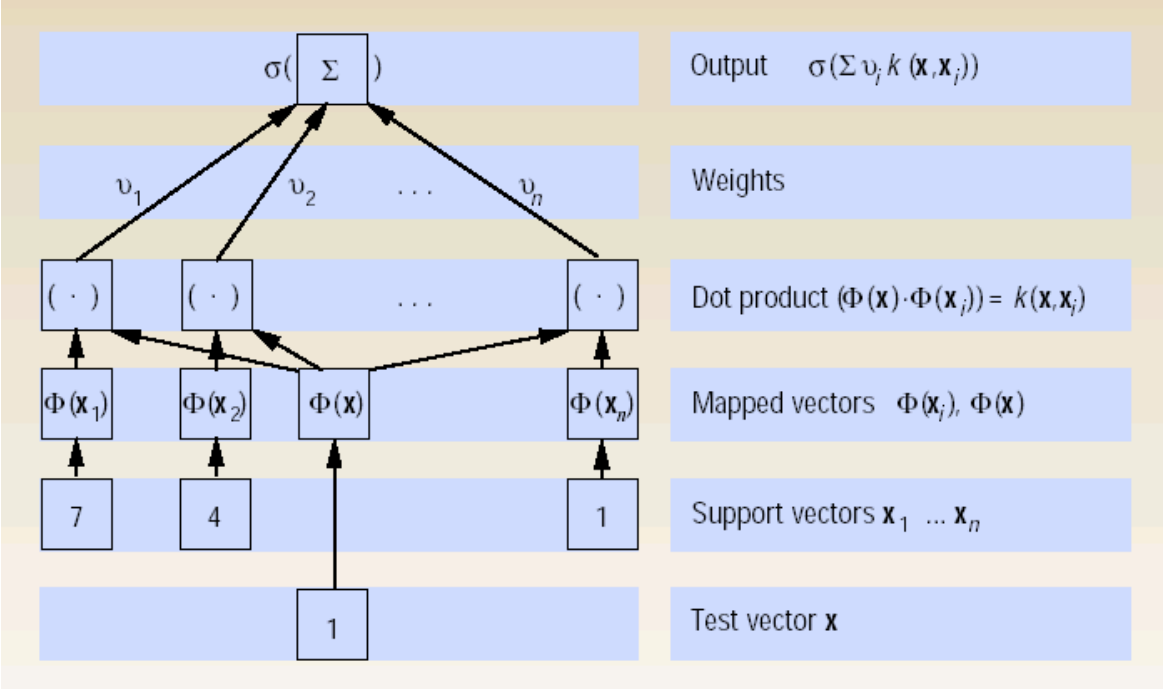
R	the set of reals
N	dimensionality of input space
F	feature space
$\boldsymbol{x}_i$	input patterns
y_i	target values (class)
l	number of training examples
$\boldsymbol{w}$	weight vector
b	constant offset (threshold)
h	VC dimension
ε	parameter of the ε-insensitive loss function
α_i	Lagrange multiplier
$\boldsymbol{\alpha}$	vector of all Lagrange multipliers
ξ_i	slack variables
‖.‖	2-norm (Euclidean distance), $\| x \| := \sqrt{(x \cdot x)}$

Figure 2. Three different views on the same dot versus cross separation problem.



Linear separation of input points (a) does not work well: a reasonably sized margin requires misclassifying one point. A better separation is permitted by nonlinear functions in input space (b), which corresponds to a linear function in a feature space (c). Input space and feature space are related by the kernel function.

Figure 3. Architecture of Support Vector method.



APPENDIX B

PUBLICATIONS RELATED TO THIS WORK

ZL Ji, LY Han, CW Yap, LZ Sun, X Chen, and YZ Chen. (2003) DART: Drug Adverse Reaction Target Database. *Drug Safety*; 26 (10): 685-690.

ZL Ji, X Chen, CJ Zhen, LX Yao, LY Han, WK Yeo, PC Chung, HS Puy, YT Tay, A Muhammad, and YZ Chen. (2003) KDBI: Kinetic Data of Bio-molecular Interactions database. *Nucleic Acids Research*; 31: 255-257.

CZ Cai, LY Han, **ZL Ji**, X Chen, YZ Chen. (2003) SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence. *Nucleic Acids Research*; 31 (13): 3692-3697.

X Chen, **ZL Ji** and YZ Chen. (2002) TTD: Therapeutic Target Database. *Nucleic Acids Research*, 30, 412 - 415.

LZ Sun, **ZL Ji**, X Chen, JF Wang, YZ Chen. (2002) ADME-AP: a database of ADME associated proteins. *Bioinformatics*; 18(12): 1699-1700.